



UNIVERSIDADE FEDERAL DO RIO GRANDE - FURG
CENTRO DE CIÊNCIAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO
CURSO DE MESTRADO EM ENGENHARIA DE COMPUTAÇÃO

Dissertação de Mestrado

**Predição da variação do teor de umidade em madeira
através do Aprendizado de Máquina**

Vítor Mendes Magalhães

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande - FURG, como requisito parcial para a obtenção do grau de Mestre em Engenharia de Computação

Orientador: Prof. Dr. Eduardo Nunes Borges
Co-orientador: Prof. Dr. Alessandro de Lima Bicho

Rio Grande, 2023

Ficha Catalográfica

M189p Magalhães, Vítor Mendes.

Predição da variação do teor de umidade em madeira através do aprendizado de máquina / Vítor Mendes Magalhães. – 2023. 89 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em Computação, Rio Grande/RS, 2023.

Orientador: Dr. Eduardo Nunes Borges.

Coorientador: Dr. Alessandro de Lima Bicho.

1. Toras de madeira 2. Teor de umidade 3. Predição
4. Aprendizado de máquina I. Borges, Eduardo Nunes II. Bicho, Alessandro de Lima III. Título.

CDU 004.8

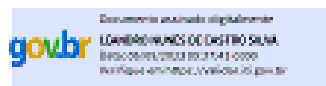
Catálogo na Fonte: Bibliotecário José Paulo dos Santos CRB 10/2344

DISSERTAÇÃO DE MESTRADO

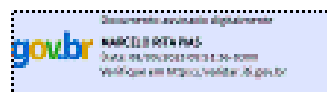
Predição da variação do teor de umidade em madeira através do
Aprendizado de Máquina

Vitor Mendes Magalhães

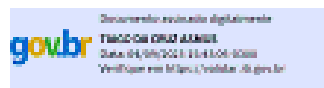
Banca examinadora:



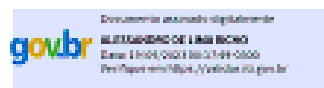
Prof. Dr. Leandro Nunes de Castro Silva (FGCU-EUA)



Prof. Dr. Marcelo Rita Pias (FURG)

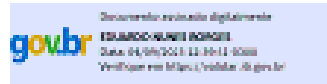


Prof. Dr. Tiago da Cruz Asmus (FURG)



Prof. Dr. Alessandro de Lima Bicho (FURG)

Coordenador



Prof. Dr. Eduardo Nunes Borges (FURG)

Orientador

AGRADECIMENTOS

Agradecer é uma dádiva. Fisicamente, ativa áreas cerebrais que liberam dopamina e ocitocina, produzindo bem-estar. Filosoficamente, torna gradativamente a percepção da vida mais positiva.

Início agradecendo à minha amada esposa, Sabrina. Seguramente, se não fosse por ela, eu não estaria escrevendo esta seção. O olhar doce e amoroso, o entendimento sobre a vida, e a capacidade de ler determinadas situações somente com a intuição, faz dela a grande responsável pelas “minhas” conquistas – então, esta conquista é tão dela quanto minha.

À minha família, que forneceu a base, o alicerce para que sobre ele fosse feita a construção que me tornou – e me torna, diariamente – o que sou.

Aos meus orientadores, Eduardo e Alessandro, que me deram um dos maiores presentes que eu poderia ter recebido em minha vida: retornar às fileiras acadêmicas e poder ter a oportunidade de estudar novamente. A sobrevivência ao mercado e a própria subsistência da família, por vezes, nos endurecem. Estudar é justamente o contrário – é um exercício diário e contínuo de humildade, pois na medida que tentamos nos aprofundar em um determinado assunto ou outro, nos damos por conta de que há um universo inteiro esperando para ser explorado, demonstrando o tamanho da nossa pequenez.

Aos meus amigos, que entenderam minha ausência por diversas vezes para assistir aulas, escrever artigos, ou para “treinar modelos”.

Aos meus colegas, e em especial, a dois deles: ao Maurício, que incontáveis vezes me ajudou *on-line* em temas que eu não tinha tanta profundidade quanto ele; e ao Jair, o colombiano mais brasileiro que eu conheço. Os dois contribuíram muito até aqui.

Por fim, mas não menos importante, aos professores do Programa de Pós-Graduação em Computação da FURG, especialmente Giancarlo, Vagner, Rafael e Hélida, que, cada um com seu jeitinho, fazem parte da minha jornada e, agora, também da minha vida.

Com todo o meu coração, muito obrigado.

*“Se és capaz de manter a tua calma quando
todo o mundo ao teu redor já a perdeu e te culpa;
de crer em ti quando estão todos duvidando,
e para esses, no entanto, achar uma desculpa.*

*Se és capaz de esperar sem te desesperares,
ou, enganado, não mentir ao mentiroso;
ou, sendo odiado, sempre ao ódio te esquivares,
e não parecer bom demais, nem pretensioso.*

*Se és capaz de pensar — sem que a isso só te atires,
de sonhar — sem fazer dos sonhos teus senhores.
Se encontrando a desgraça e o triunfo, conseguires
tratar da mesma forma a esses dois impostores.*

*Se és capaz de sofrer a dor de ver mudadas
em armadilhas as verdades que disseste,
e as coisas, por que deste a vida, estraçalhadas,
e refazê-las com o bem pouco que te reste.*

*Se és capaz de arriscar numa única parada
tudo quanto ganhaste em toda a tua vida,
e perder! E, ao perder, sem nunca dizer nada,
resignado, tornar ao ponto de partida.*

*De forçar coração, nervos, músculos, tudo
a dar seja o que for que neles ainda existe;
e a persistir assim quando, exaustos, contudo
resta a vontade em ti que ainda ordena: “Persiste!”;*

*Se és capaz de, entre a plebe, não te corromperes
e, entre reis, não perder a naturalidade;
e de amigos, quer bons, quer maus, te defenderes,
se a todos podes ser de alguma utilidade.*

*Se és capaz de dar, segundo por segundo,
ao minuto fatal todo o valor e brilho;
tua é a Terra com tudo o que existe no mundo
e — mais ainda — serás um homem, meu filho!”*

— RUDYARD KIPLING

RESUMO

MAGALHÃES, Vítor Mendes. **Predição da variação do teor de umidade em madeira através do Aprendizado de Máquina**. 2023. 88 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande - FURG, Rio Grande.

A madeira é uma das matérias-primas mais utilizadas para a fabricação de diversos produtos ao redor do mundo. Em se tratando de processos produtivos e industriais, pelo fato de a madeira ser oriunda de árvores – um recurso natural – é preciso otimizar ao máximo sua utilização. Um ponto-chave para essa otimização é o controle da variação de umidade presente na madeira, pois, na medida em que as toras de madeira bruta são armazenadas ao ar livre, elas começam a perder água de forma gradativa. Tal perda de água é refletida diretamente não somente no peso, mas também no surgimento de defeitos, como empenos e rachaduras, que prejudicam de forma peremptória os processos produtivos que as toras serão submetidas, causando perdas irreparáveis de matéria-prima. Dado o contexto, esta dissertação de mestrado desenvolve um método baseado em *Machine Learning* para prever a variação do peso de toras de madeira causada pela perda de umidade, sem que seja necessário adquirir equipamentos ou aumentar o custo de mão-de-obra, utilizando dados operacionais e meteorológicos que, em via de regra, as empresas do setor já têm à disposição. São ajustados diversos modelos, com diferentes algoritmos, configurações de hiperparâmetros e distribuições dos dados de treinamento. Esses modelos foram submetidos a diferentes avaliações. Como avaliação empírica, a métrica escolhida foi a medida F balanceada, ponderada pela quantidade de instâncias por classe, referenciada no texto como μF_1 . Os resultados dos experimentos mostraram que o modelo desenvolvido utilizando o classificador *Random Forest*, classes balanceadas pelo algoritmo SMOTE e hiperparâmetros ajustados com o *best tuning* obteve o melhor desempenho de classificação. Na análise estatística, foram empregados os testes *Align-Friedman* e *post hoc* de *Holm*. Considerando o modelo baseado no *Random Forest* como método de controle, os resultados demonstraram que os algoritmos *Random Forest* e *Gradient Boosting* podem ser considerados estatisticamente equivalentes na maioria dos cenários testados. Dessa forma, sugere-se para utilização industrial o modelo baseado no *Random Forest*, possibilitando, assim, um melhor aproveitamento das toras de madeira nos processos produtivos e auxiliando a tomada de decisão da cadeia industrial na qual a madeira está inserida.

Palavras-chave: Toras de madeira, teor de umidade, predição, aprendizado de máquina.

ABSTRACT

MAGALHÃES, Vítor Mendes. **Prediction of wood moisture content variation with Machine Learning**. 2023. 88 f. Dissertação (Mestrado) – Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande - FURG, Rio Grande.

Wood is one of the most used raw materials for the manufacture of various products around the world. When it comes to production and industrial processes, because wood comes from trees – a natural resource – it is necessary to optimize its use as much as possible. A key point for this optimization is the control of the variation of moisture content present in the wood, because as the raw wood logs are stored in the open air, they gradually start to lose water. Such water loss is directly reflected not only in weight, but also in the appearance of defects, such as warping and cracks, which peremptorily impair the production processes that the logs will be subjected to, causing irreparable losses of raw material. Given the context, this master's thesis develops a method based on *Machine Learning* to predict the variation in the weight of wood logs caused by moisture content loss, without the need to purchase equipment or increase the cost of labor, using operational and meteorological data that, as a rule, companies in the sector already have at their disposal. Several models are adjusted, with different algorithms, hyperparameter configurations and training data distributions. These models were submitted to different evaluations. As an empirical evaluation, the metric chosen was the balanced F measure, weighted by the number of instances per class, referred in the text as μF_1 . The results of the experiments showed that the model developed using the *Random Forest* classifier, classes balanced by the SMOTE algorithm and hyperparameters adjusted with *best tuning* obtained the best classification performance. In the statistical analysis, *Align-Friedman* and *post hoc* tests by *Holm* were used. Considering the model based on *Random Forest* as a control method, the results showed that the algorithms *Random Forest* and *Gradient Boosting* can be considered statistically equivalent in most of the scenarios tested. Thus, a model based on *Random Forest* is suggested for industrial use, thus enabling better use of wood logs in production processes and helping decision-making in the industrial chain in which wood is inserted.

Keywords: wood logs, moisture content, prediction, machine learning.

LISTA DE FIGURAS

Figura 1	Exemplo de toras de madeira armazenadas em uma pilha em um pátio de armazenagem. Fonte: o autor.	13
Figura 2	Alguns subprodutos das toras de madeira: cavacos (à esquerda) e laminados (à direita). Fonte: [61]	14
Figura 3	Pilha de toras de eucalipto, com destaque para as rachaduras das seções transversais. Fonte: o autor.	15
Figura 4	Adaptação do processo de descoberta de conhecimento em bancos de dados. [25]	44
Figura 5	Adaptação da sintetização do processo de descoberta de conhecimento em bancos de dados. [17]	44
Figura 6	Distribuições das temperaturas durante o período observado de armazenamento das toras de madeira.	49
Figura 7	Distribuições de precipitação total, pressão atmosférica média, umidade relativa mínima e umidade relativa média durante o período observado de armazenamento das toras de madeira.	50
Figura 8	Distribuições de velocidade média do vento, e das rajadas máximas médias do vento durante o período observado de armazenamento das toras de madeira.	50
Figura 9	Distribuições da quantidade de dias em que as toras estiveram armazenadas nas pilhas durante o período observado.	51
Figura 10	À esquerda, o <i>boxplot</i> com os percentuais de perda de peso. À direita, o histograma de frequência das instâncias com diferentes percentuais de perda de peso.	53
Figura 11	Comparação de desempenho dos algoritmos em diferentes configurações de treinamento.	64
Figura 12	Distribuição da μF_1 considerando dados desbalanceados sem ajuste de hiperparâmetros.	65
Figura 13	Distribuição da μF_1 considerando dados desbalanceados e hiperparâmetros ajustados.	66
Figura 14	Distribuição da μF_1 considerando dados balanceados sem ajuste de hiperparâmetros.	66
Figura 15	Distribuição da μF_1 considerando dados balanceados e hiperparâmetros ajustados.	67
Figura 16	Métricas F1 de cada modelo em cada configuração dos dados.	67
Figura 17	Importância de cada atributo para o modelo analisado, considerando os dez atributos mais importantes.	75

LISTA DE TABELAS

Tabela 1	Matriz de confusão binária (duas classes).	21
Tabela 2	Exemplo de matriz de confusão considerando um problema com quatro classes.	22
Tabela 3	Linha do tempo e características dos estudos	41
Tabela 4	Linha do tempo e principais contribuições das pesquisas	42
Tabela 5	Detalhamento dos atributos presentes no conjunto final de dados. . .	48
Tabela 6	Divisão de intervalos de perda de peso.	55
Tabela 7	Balanceamento entre as classes do conjunto de dados.	55
Tabela 8	Espaço de busca dos hiperparâmetros de cada algoritmo utilizado na avaliação experimental.	62
Tabela 9	O <i>best tuning</i> de cada modelo.	62
Tabela 10	Desempenho sobre os dados de teste de todos os modelos treinados considerando a média ponderada da F1.	63
Tabela 11	<i>Ranking</i> após os testes estatísticos sobre o conjunto 1.	69
Tabela 12	<i>Ranking</i> após os testes estatísticos sobre o conjunto 2.	70
Tabela 13	<i>Ranking</i> após os testes estatísticos sobre o conjunto 3.	70
Tabela 14	<i>Ranking</i> após os testes estatísticos sobre o conjunto 4.	71
Tabela 15	Melhor configuração de hiperparâmetros do melhor modelo encontrado. .	71
Tabela 16	Matriz de confusão do algoritmo <i>Random Forest</i> , sobre dados balanceados e hiperparâmetros ajustados através do seu <i>best tuning</i> , na <i>seed</i> 100000.	72
Tabela 17	Métricas do algoritmo <i>Random Forest</i> para cada classe, sobre dados balanceados e hiperparâmetros ajustados através do seu <i>best tuning</i> , na <i>seed</i> 100000.	73
Tabela 18	Casos de falha do modelo.	78

LISTA DE ABREVIATURAS E SIGLAS

AB	AdaBoost
ANN	Artificial Neural Networks
DT	Decision Trees
GB	Gradient Boosting
IA	Inteligência Artificial
IIR	Infinite Impulse Response
INMET	Instituto Nacional de Meteorologia
KDD	Knowledge Discovery in Databases
LR	Logistic Regression
LS-SVM	Least Squares Support Vector Machines
MAPA	Ministério da Agricultura, Pecuária e Abastecimento do Brasil
ML	Machine Learning
MLP	Multi-Layer Perceptrons
MRAE	Mean Relative Absolute Error
NB	Gaussian Naive Bayes
RF	Random Forest
SMOTE	Synthetic Minority Over-Sampling Technique
SVM	Support Vector Machines
SVR	Support Vector Regression

SUMÁRIO

1	Introdução	12
1.1	Objetivo do trabalho	15
1.2	Organização do trabalho	16
2	Fundamentação teórica	18
2.1	Aprendizado de Máquina	18
2.2	Classificação	20
2.3	Avaliação de classificadores	21
2.4	Algoritmos de classificação	23
2.5	Lidando com o desbalanceamento de classes	28
3	Trabalhos relacionados	30
3.1	Modelos baseados em Aprendizado de Máquina	30
3.2	Modelos analíticos, estatísticos, ou baseados em generalizações	33
3.3	Outros trabalhos correlatos	37
3.4	Comparativo dos estudos relacionados no levantamento bibliográfico	40
4	Metodologia	43
4.1	Conjunto de dados	45
4.1.1	Obtenção dos dados	46
4.1.2	Descrição dos atributos	47
4.1.3	Estatística descritiva dos dados	47
4.1.4	Pré-processamento dos dados	51
4.2	Mineração dos dados	56
5	Resultados	63
5.1	Avaliação empírica	63
5.2	Análise estatística	67
5.3	Melhor modelo	72
5.3.1	Análise dos casos de falha	73
5.4	Ética na utilização do modelo	77
6	Conclusão	79
	Referências	81

1 INTRODUÇÃO

O crescimento econômico exponencial da China e sua consequente transformação estrutural de uma sociedade essencialmente rural para uma sociedade urbano-industrial inevitavelmente produziria demandas por recursos naturais de toda ordem. Nos últimos vinte anos, o Brasil tornou-se o maior fornecedor de produtos agrícolas para o mercado chinês. As exportações para a China aceleraram e aprofundaram as mudanças tecnológicas e organizacionais na agricultura brasileira – que estão na base da competitividade do agro-negócio brasileiro [88].

Em se tratando de madeira, dado o seu acesso limitado a recursos naturais, a China tem recorrido cada vez mais às compras externas para a produção de suas fábricas locais. Esse fato levou o Brasil a dobrar as exportações de toras de madeira entre 2017 e 2018. Segundo levantamento da consultoria *Forest2Market*, com base em dados de comércio exterior do governo brasileiro, a exportação de toras de eucalipto totalizou 234 mil toneladas em 2018, 122% acima do registrado em 2017. Desse total, 89% foi exportado para China [76].

Desde as últimas duas décadas, a indústria florestal e madeireira da China se desenvolve de forma extremamente acelerada. Este progresso acabou tornando a China o maior *player* não apenas no mercado global de madeira bruta, mas também no mercado de produtos intermediários e produtos acabados de madeira, resultando na China sendo o maior exportador de produtos de madeira com valor agregado [6].

É necessário também somar a demanda nacional por este insumo, já estudada para médio e longo prazo. Entre 1988 e 2007, estudos já apresentavam resultados indicando que as variáveis preço da madeira em tora e capacidade instalada da indústria de celulose explicavam a demanda nacional [5].

Desde o passado o Brasil prevê o aumento da demanda de madeira. A partir da década de 60, diversos incentivos fiscais foram criados para a produção de madeira através do plantio de espécies de rápido crescimento – especialmente com os gêneros *Pinus* e *Eucalipto*, especialmente de 1966 a 1988 [42].

Nesse sentido, se faz importante observar que o suprimento para as indústrias de madeira pelas florestas plantadas, em substituição à extração de madeira de florestas de ori-



Figura 1: Exemplo de toras de madeira armazenadas em uma pilha em um pátio de armazenagem. Fonte: o autor.

gem nativa, é crescente. De acordo com dados do IBGE de 2014, a participação da madeira plantada em relação ao total consumido passou de cerca de 20% em 1990 para mais de 70% em 2013 [42].

Ainda que avanços estejam sendo conquistados de forma continuada, o Brasil historicamente sempre teve dificuldades para promover a expansão das florestas plantadas de forma duradoura e, ao mesmo tempo, controlar o desmatamento. Para agricultores que não estão habituados à cadeia produtiva florestal, algumas características podem reduzir drasticamente sua atratividade, como por exemplo o retorno acontecer apenas após muitos anos do momento do cultivo.

Assim, o desmatamento tem se tornado uma crescente preocupação, principalmente em países em desenvolvimento, como o Brasil, nos quais o controle das áreas florestais é, muitas vezes, realizado de forma precária. De forma generalizada, o desmatamento geralmente possui repercussões globais. Perdas em larga escala de áreas florestais têm implicado em mudanças na própria oferta de madeira, no equilíbrio hidrológico, nos recursos genéticos e em ciclos de carbono [2].

Independente da destinação, seja para uso doméstico ou para exportação, muitos produtos acabados que possuem a madeira como matéria-prima passam pela mesma etapa: o armazenamento das toras de madeira em pilhas. A Figura 1 apresenta uma fotografia de um pátio de toras de uma empresa do setor com uma pilha em destaque.

Toras de madeira de árvores recém-abatidas apresentam elevado teor de água. Este teor vai se reduzindo, lentamente, na medida em que as toras são expostas às condições ambientais de determinado local [69]. Alterações do teor de umidade da madeira levam a alterações de praticamente todas as suas propriedades físicas e mecânicas. É necessário que tais alterações sejam entendidas e levadas em consideração para o consequente uso

da madeira [29].

Durante o armazenamento, a madeira seca e seu peso diminui com a perda de umidade [85]. Por isso, o tempo de armazenamento das toras de madeira em pilhas é um fator determinante para a qualidade do produto acabado, pois influencia diretamente no aceite, caso a madeira tenha por destino a fabricação de cavaco [48] ou laminados [40]. A Figura 2 ilustra estes dois tipos de subprodutos das toras.



Figura 2: Alguns subprodutos das toras de madeira: cavacos (à esquerda) e laminados (à direita). Fonte: [61]

Sendo assim, torna-se facilmente perceptível que o tempo de armazenamento das toras pode resultar em perdas causadas por rachaduras – o que diminui consideravelmente a produtividade de indústrias de laminados de madeira, por exemplo. Por esta razão, o teor de umidade é visto com um fator que deve ser observado e mantido dentro dos padrões [24]. A Figura 3 apresenta exemplos de seções transversais de toras de madeira, armazenadas em pilhas, com diferentes condições de rachadura.

Sabendo-se que a madeira pode absorver de 25 a 30% de seu peso em água [19], depreende-se que as rachaduras são responsáveis pelo grande volume de perdas na madeira durante o armazenamento das toras [30].

Para diminuir a incidência e até evitar rachaduras, existem alguns dispositivos que são utilizados pela indústria na cadeia florestal. O mais utilizado é o chamado *anti-racha*, que consiste em duas placas de metal que são instaladas nas extremidades das toras, recobrendo uma área mínima de 70% das seções transversais [78]. Tais placas têm por finalidade diminuir a possibilidade de incidência de rachaduras nas toras, evitando prováveis perdas no processo industrial caso não fossem fixadas – especialmente no processo de laminação.

Para isso, se faz imperativo prever o momento exato no qual a variação de umidade atinge o nível suficiente para que as toras comecem a rachar. Dessa forma, medidas que dificultam e até mesmo impedem o surgimento de rachaduras poderão ser tomadas – como a fixação de placas anti-rachas, por exemplo. De igual modo, a cadeia de transporte também poderá ser diretamente beneficiada, pois com a utilização de um método de predição de variação da umidade, sabendo-se que existe relação direta entre a umidade



Figura 3: Pilha de toras de eucalipto, com destaque para as rachaduras das seções transversais. Fonte: o autor.

e o peso da madeira [65], todo o transporte poderia ser otimizado. Por último, mas não menos importante, em se tratando do processo de laminação de toras de madeira, a produtividade das toras de madeira será consideravelmente aumentada, ou seja, o suprimento das mesmas florestas atuais à indústria poderá ser igualmente otimizado.

1.1 Objetivo do trabalho

Partindo-se da premissa de que existe relação direta entre o teor de umidade presente nas toras de madeira e o peso da madeira armazenada em toras [65], o objetivo geral deste trabalho é desenvolver um modelo efetivo de predição da variação de umidade em toras de madeira utilizando diversos atributos preditores, como informações biológicas, florestais e morfológicos das toras de madeira, além de informações geográficas e meteorológicas do seu ambiente de armazenamento. Como objetivos específicos, destacam-se ainda:

1. investigar as relações entre condições meteorológicas e a variação de umidade das toras;
2. caracterizar diferentes perfis de variação de umidade para cada espécie analisada;
3. implantar o modelo preditivo ajustado com dados históricos e avaliar sua eficácia em campo, ou seja, em ambiente relevante ou operacional;
4. analisar os casos de falha do melhor modelo ajustado.

Para se atingir estes objetivos, as seguintes atividades são necessárias: analisar a literatura científica relevante, através de uma revisão bibliográfica, para melhor compreensão dos métodos utilizados atualmente pelo mercado internacional de madeira; obter um conjunto de dados, contendo os dados específicos das pilhas de toras de madeira, além dos dados meteorológicos; pré-processar o conjunto de dados, visando remover, padronizar e melhor compreender os atributos; analisar a importância dos atributos preditores e a

forma com que se relacionam com a variação do peso das toras, utilizando métodos de *feature selection*; delimitar possíveis classes do problema, após análise técnica com um especialista; definir a abordagem de validação dos modelos; aplicar diferentes algoritmos preditivos baseados em *Machine Learning* para o desenvolvimento dos modelos; analisar estatisticamente os resultados dos diferentes modelos desenvolvidos para a resolução do problema proposto; e, por fim, analisar os resultados e sugerir a utilização do melhor modelo em testes de campo.

1.2 Organização do trabalho

Este trabalho está organizado na seguinte forma: no Capítulo 2, a fundamentação teórica do presente trabalho é apresentada, com ênfase na abordagem sobre *Machine Learning* (Seção 2.1) utilizando métodos de classificação (Seção 2.2).

Após, no Capítulo 3, são discutidos os trabalhos relacionados com o presente objeto de pesquisa. Estão descritos estudos relevantes com modelos utilizando métodos de *Machine Learning* (Seção 3.1), modelos utilizando bases estatísticas, analíticas e generalistas (Seção 3.2) e ainda outros estudos que não apresentam de forma direta modelos preditivos, mas, ou estão relacionados ao problema de pesquisa, ou estão relacionados à solução proposta (Seção 3.3). Na Seção 3.4 os trabalhos são resumidos em uma tabela comparativa, citando os estudos evidenciados, bem como suas principais contribuições.

O Capítulo 4 trata exclusivamente da solução proposta para o problema de pesquisa do presente trabalho, apresentando a metodologia adotada. São destacados os atributos preditores e o atributo-alvo – o que se objetiva prever.

Inicialmente, é apresentada uma explanação detalhada sobre o conjunto de dados utilizado (Seção 4.1), bem como sobre todos os procedimentos adotados para a obtenção dos mesmos (Subseção 4.1.1), além do primeiro pré-processamento dos dados meteorológicos. A seguir, uma tabela descreve detalhadamente os atributos presentes no conjunto de dados (Subseção 4.1.2), evidenciando os intervalos e a média dos atributos numéricos, e a moda dos atributos categóricos. Adiante, é apresentada uma análise descritiva (Subseção 4.1.3), de forma a aprofundar o conhecimento sobre os mesmos dados e elucidar o ambiente no qual o problema está inserido. A seção referente ao conjunto de dados é finalizada com a etapa de pré-processamento (Subseção 4.1.4), que tem por objetivo demonstrar todas as ações e procedimentos realizados para que a etapa seguinte – a mineração dos dados – fosse melhor aproveitada.

A Seção 4.2 foca na extração dos padrões ocultos presentes nos dados e na produção do conhecimento, de forma a sugerir para a indústria do setor uma solução do problema de pesquisa em questão, utilizando técnicas de *Machine Learning*. Nesta seção são apresentados os diferentes algoritmos que foram utilizados para o desenvolvimento dos modelos de predição.

Na Capítulo 5, os resultados obtidos na etapa de mineração de dados são compilados, formatados, analisados de diferentes formas, e discutidos, com o objetivo de validar a solução proposta para resolver o presente problema de pesquisa industrial. Com base em análises e testes estatísticos (Seção 5.2), a capacidade de generalização dos modelos propostos é demonstrada.

Finalizando esta dissertação, o Capítulo 6 trata da conclusão do estudo realizado, bem como das sugestões para trabalhos futuros, objetivando a continuidade no desenvolvimento de soluções distintas – e cada vez melhores – para o tema em questão.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção define os conceitos fundamentais sobre o objeto da presente pesquisa, evidenciando terminologias, abordando conceitos e nivelando conhecimentos.

2.1 Aprendizado de Máquina

Pelo fato de o objetivo desta pesquisa ser, de fato, desenvolver modelos que consigam prever a variação do teor de umidade – através do peso – de toras de madeira em processo de secagem natural utilizando *Machine Learning*, faz-se necessário aprofundar o conhecimento sobre o tema.

Originalmente, o Aprendizado de Máquina, ou *Machine Learning*, pode ser entendido como um programa de computador que aprende com a experiência no que diz respeito a alguma classe de tarefas, avaliando se o seu desempenho ao realizar estas tarefas melhora com a experiência [54].

O Aprendizado de Máquina pode ser entendido como um subcampo da Inteligência Artificial (IA) [71] que tem a capacidade de construir modelos matemáticos com a finalidade de fazer previsões, ou tomar decisões, sem precisar ser explicitamente programado para isso [95], aprendendo com o ambiente ao redor [22]. Os algoritmos de aprendizado de máquina são, de certa forma, “auto-codificados”, no sentido de que eles podem se alterar ou se adaptar de forma automática através da experiência (repetição), tornando-se cada vez melhores na realização de uma determinada tarefa.

O processo de aprendizado de máquina pode se utilizar de diferentes algoritmos para resolver problemas. Usualmente, salienta-se que não existe um mesmo algoritmo que resolva melhor todas as tarefas [1]. Por isso, incorporar o conhecimento específico do problema no comportamento do algoritmo se torna extremamente importante [92], assim como compreender as limitações dos algoritmos, preferencialmente utilizando-se de metodologias que permitam avaliar os conceitos induzidos pelos mesmos na resolução de determinados problemas. O algoritmo, então, que deve ser abordado, depende diretamente do tipo de problema que se deseja resolver, do número de variáveis e do modelo que melhor se adapta [50].

A grande área do aprendizado de máquina pode ser subdividida em dois principais – e diferentes – tipos de aprendizado: aprendizado supervisionado e aprendizado não-supervisionado [11]. Embora existam outros tipos de aprendizado de máquina, como o aprendizado por reforço e o aprendizado semi-supervisionado, por exemplo, alguns autores ainda classificam os demais tipos de aprendizado como partícipes de algum dos conjuntos principais. Como o objetivo deste trabalho não é comparar diferentes tipos de aprendizado de máquina, será utilizada a definição fundamentalista.

O aprendizado *supervisionado* reflete a capacidade que um determinado algoritmo tem de generalizar conhecimento a partir de dados disponibilizados com uma etiqueta (rótulo), ou seja, com um atributo-alvo, para que ele possa ser utilizado para prever novos casos que ainda não foram rotulados [3]. O aprendizado *não-supervisionado* se refere a aplicações sobre dados que não foram classificados ou categorizados, ou seja, não possuem etiqueta, envolvendo o reconhecimento de padrões sem a interferência de um atributo-alvo. Dessa forma, no aprendizado não-supervisionado, todos os atributos presentes no conjunto de dados são utilizados como entradas do modelo [11], fazendo com que os algoritmos aprendam os relacionamentos intrínsecos entre os dados disponíveis, agrupando ou associando os casos de acordo com algumas regras determinadas. Quando somente poucos dados estão disponíveis de forma rotulada, o aprendizado ainda pode ser classificado, segundo alguns autores, como *semi-supervisionado* – embora a literatura relevante ainda não classifique como uma terceira categoria de aprendizado.

De acordo com a natureza da tarefa de aprendizado, preditiva ou descritiva, os algoritmos podem ser categorizados da seguinte forma: enquanto os algoritmos de aprendizado não-supervisionado são subdivididos em *agrupamento* e *associação*, os algoritmos de aprendizado supervisionado são subdivididos em algoritmos de *classificação* e algoritmos de *regressão* [4].

Neste estudo, o atributo-alvo presente no conjunto de dados foi rotulado. Portanto, é possível desenvolver um modelo, utilizando a abordagem de aprendizado de máquina, para resolver o problema. Sendo assim, trata-se de um problema de *aprendizado de máquina supervisionado*.

Em se tratando de aprendizado supervisionado, ao verificar o conjunto de dados e sua relação com o tipo de problema que deve ser resolvido, é necessário analisar o atributo-alvo através de duas distintas perspectivas: quando o tipo de dados do atributo-alvo é categórico ou ordinal, trata-se de um problema de *classificação*; já quando o tipo de dados do atributo-alvo é numérico e contínuo, trata-se de um problema de *regressão* [33].

Em nova referência, sendo o atributo-alvo do presente estudo um tipo de dado numérico e contínuo – peso de toras de madeira, depreendia-se que, além de ser um problema de aprendizado supervisionado, pudesse tratar-se de um problema de *regressão*. Entretanto, buscando uma maior assertividade e, por consequência, uma maior aderência da solução proposta com o ambiente industrial, optou-se por criar outro atributo-alvo, de

forma dependente, estabelecendo diferentes classes de intervalos de variação de peso das toras de madeira. Assim, o problema deixará de ser tratado como *regressão* e passará a ser um problema de *classificação*.

Ao ser desenvolvido um novo modelo utilizando *Machine Learning*, percorrer algumas etapas se faz necessário. A primeira etapa é a coleta e seleção dos dados. É importante ter sempre em mente que é somente a partir dos dados que os padrões – até então ocultos – podem ser “aprendidos” pelos algoritmos. Como a grande maioria dos algoritmos de *Machine Learning* requer um grande montante de dados para possibilitar a inferência com assertividade significativa, a etapa de coleta dos dados pode ser desafiadora.

A etapa seguinte é o pré-processamento. É a etapa na qual ocorrem os tratamentos de possíveis ruídos nos dados, que podem afetar de forma significativa a qualidade dos modelos que serão treinados *a posteriori*. Ao final da etapa de pré-processamento, é possível – mas, dependendo do problema, não necessário – selecionar as características dos dados presentes que são mais importantes para o desenvolvimento dos modelos [12].

Somente após estas etapas que o algoritmo é executado sobre os dados – já pré-processados e com características selecionadas. Esta execução que ajusta um modelo preditivo é chamada de *treinamento*. Neste processo, são fornecidos os dados de entrada e os rótulos – que são os resultados desejados – e o algoritmo então pode ser configurado para que possa não apenas produzir o resultado desejado, mas também generalizar, ou seja, ter a mesma capacidade preditiva quando o modelo é submetido a novos dados ainda não vistos anteriormente.

A etapa de treinamento é justamente o ponto no qual acontece o *aprendizado* da máquina. E se faz importante salientar que o algoritmo não precisa se limitar apenas à sua configuração inicial: assim como ocorre com os humanos, os algoritmos podem praticar o processo de aprendizagem *ad infinitum*, na medida em que vão processando novos dados de entrada e aprendendo com os seus próprios erros. Portanto, sistemas que utilizam algoritmos de aprendizado de máquina são programas de computador que tomam decisões baseadas em experiências acumuladas através da solução bem sucedida de problemas anteriores [55], como uma espécie de retroalimentação.

2.2 Classificação

Na área de *Machine Learning*, os modelos de classificação são muito estudados e, possivelmente, os que possuem maior relevância prática na resolução de problemas reais.

A classificação é uma abordagem de aprendizado de máquina supervisionado que é utilizada para prever a associação de novas instâncias de dados a determinadas categorias, também chamadas de *classes*. Assim, a classificação consiste na construção de um procedimento que é capaz de prever – ou *classificar* – com sucesso a classe verdadeira

de uma determinada instância, que ainda não é conhecida, tendo por base um conjunto de casos em que as classes já são conhecidas.

Comparando os modelos de classificação com os modelos de regressão, pode-se afirmar que enquanto os modelos de regressão mapeiam o espaço de entrada em um domínio de valor real, os modelos de classificação mapeiam o espaço de entrada em classes predefinidas [58].

Matematicamente, um classificador pode ser definido como uma função f que mapeia vetores de características $x \in X$ como entrada em rótulos de classe $y \in \{1, \dots, C\}$ como saída, onde X é o universo de características e C é o número de classes considerado. Via de regra, assume-se que $X = \mathbb{R}^D$ ou $X = \{0, 1\}^D$, ou seja, que o vetor de características é um vetor de D números reais ou D bits, mas, em geral, pode-se misturar características tanto discretas quanto contínuas. Também se assume que os rótulos (classes) são desordenados (categóricos) e mutuamente exclusivos. Ao permitir que uma determinada entrada pertença a mais de uma classe, isso é chamado de um *problema multi-classe*.

Um modelo de classificação tem seu desempenho mensurado com base na sua taxa de acerto e/ou de erro. As estimativas dessas taxas não somente fornecem uma base objetiva, como também podem ser ferramentas poderosas para a tomada de decisão sobre determinados algoritmos de classificação [9]. A seguir são apresentadas algumas métricas de avaliação de classificadores utilizadas na avaliação experimental realizada neste trabalho.

2.3 Avaliação de classificadores

As métricas utilizadas na avaliação de modelos preditivos com atributo-alvo categórico são normalmente baseadas na matriz de confusão [83]. Esta matriz contém todas as informações relevantes sobre o desempenho das predições geradas pelo modelo de classificação. A matriz de confusão cruza, para cada classe, a quantidade de instâncias reais (*reference*) e preditas pelo modelo no conjunto de dados avaliado. Via de regra, as colunas representam as predições. As possíveis classes são listadas na mesma ordem nas linhas e nas colunas. Desta forma, as instâncias classificadas corretamente ficam localizadas na diagonal principal da matriz.

Na Tabela 1 é possível verificar visualmente a configuração-padrão e a disposição de uma matriz de confusão binária (duas classes), onde vp simbolizam os verdadeiros positivos, vn os verdadeiros negativos, fp os falsos positivos e fn os falsos negativos.

Tabela 1: Matriz de confusão binária (duas classes).

		Predito	
		Positivo	Negativo
Real	Positivo	vp	fn
	Negativo	fp	vn

Diversas métricas podem ser obtidas partindo da matriz de confusão, tendo quatro

delas como principais: a acurácia (*accuracy*), a precisão (*precision*), a revocação (*recall*) e a medida F balanceada entre precisão e revocação (*F1-score*).

A mais popular entre as métricas – a acurácia – é o percentual de instâncias corretamente classificadas, independente da classe de dados. Matematicamente, pode ser entendida através da razão:

$$accuracy = \frac{vp + vn}{vp + vn + fp + fn}$$

Na matriz de confusão exemplificada na Tabela 2, a acurácia do modelo é de 0,612, ou seja, 61,2%.

Tabela 2: Exemplo de matriz de confusão considerando um problema com quatro classes.

		Predito			
		A	B	C	D
Real	A	41	10	2	1
	B	9	16	6	0
	C	1	3	3	0
	D	2	1	3	0

A precisão, que simboliza a habilidade que um modelo tem de não classificar como positiva uma instância que é negativa, pode ser entendida por:

$$precision = \frac{vp}{(vp + fp)}$$

A revocação simboliza a habilidade de um modelo identificar instâncias positivas. Pode ser entendida por:

$$recall = \frac{vp}{(vp + fn)}$$

Por último, a F1 é a média harmônica entre a precisão e a revocação:

$$F1 = \frac{2 \times (precision \times recall)}{(precision + recall)}$$

Considerando o mesmo exemplo da Tabela 2, a precisão da classe B é 53,3%, a revocação da classe A é 75,9% e a F1 da classe C é 30%.

Com exceção da acurácia, que é uma medida global, as demais métricas de avaliação são aplicadas para cada classe dos dados. No entanto, é possível obter uma medida global através de uma agregação, representada por diferentes médias: *macro* e *micro* (μ). A primeira é a média aritmética simples entre as medidas de todas as classes, ou seja, caso exista desbalanceamento de classes nos dados, esta média pode esconder o real desempenho do modelo. A μ -média, ou *avgweighted* representa a média ponderada da métrica,

dado o número de instâncias de cada classe. A utilização desta medida é recomendada quando o conjunto de dados é desbalanceado – como é o caso deste trabalho.

Em se tratando de conjuntos de dados desbalanceados, a acurácia não é uma métrica representativa. Um bom modelo de classificação deve encontrar o máximo de exemplos positivos (alta revocação) sem sacrificar a precisão [10], ou seja, retornando a menor quantidade possível de falsos positivos (alta precisão).

Classes desbalanceadas podem comprometer fortemente o processo de descoberta de conhecimento, pois o modelo tende a focar o aprendizado na classe majoritária (predominante) e ignorar as classes minoritárias (raras). Então, além da geração do modelo de classificação ser afetada pela distribuição assimétrica das classes, o próprio processo de avaliação do modelo também é afetado [52]. Bastaria um modelo classificar como verdadeiros positivos todas as instâncias da classe majoritária, ainda que ignorasse as classes minoritárias, para fornecer o que seria considerada uma boa acurácia. Entretanto, é comum que principal interesse nas aplicações descoberta de conhecimento esteja, de fato, no aprendizado justamente da classe minoritária¹. Diferentes algoritmos para lidar com o desbalanceamento de classes são discutidos a Seção 2.5.

2.4 Algoritmos de classificação

Dentre os algoritmos de classificação, em se tratando de métodos de *Machine Learning*, se faz importante destacar alguns, como C4.5 [63], *Random Forest* (RF) [13], *Logistic Regression* (LR) [46], *Support Vector Machines* (SVM) [80], *Naive Bayes* (NB) [57], *AdaBoost* (AB) [34], *Gradient Boosting* (GB) [26], além da classe de métodos que se utiliza de *Artificial Neural Networks* (ANN) [28].

A abordagem utilizada pelo algoritmo de floresta aleatória (*Random Forest* – RF) é diferente da abordagem dos algoritmos baseados somente em árvores de decisão, tais como ID3 [64] e C4.5 [63]. Primeiramente, RF realiza o processo conhecido como *bootstrapping*: a extração de novos conjuntos de dados de forma aleatória, do conjunto original. Cada subconjunto contém o mesmo número de instâncias do conjunto original, diferindo apenas nas características, selecionadas igualmente de forma aleatória, como um processo simplificado e organizado de amostragem. Tal etapa se faz extremamente importante para este algoritmo, pois irá garantir que diferentes dados serão utilizados para a criação de cada árvore na floresta (tornando o algoritmo menos sensível aos dados originais), além da seleção aleatória de características ajudar a reduzir a correlação entre as árvores da floresta. Estas propriedades do algoritmo evitam sobreajuste do modelo².

¹ Considere o exemplo hipotético da predição de fraudes em compras por cartões de crédito em que, num conjunto de 100 instâncias, 95 delas sejam legítimas e 5 sejam fraudulentas; um modelo que classificasse somente a classe majoritária e ignorasse as transações fraudulentas teria uma acurácia de 95%. Entretanto, o principal objetivo do modelo – o *aprendizado* – estaria justamente no oposto: em aprender a identificar as compras fraudulentas.

² Sobreajuste, ou *overfitting*, é um termo utilizado para descrever quando um modelo estatístico se ajusta

O passo seguinte é o treinamento de diferentes árvores de decisão, uma para cada subconjunto. Ao conjunto destas árvores, chama-se *floresta* – daí uma parte do nome do algoritmo. Cada árvore de decisão é capaz de realizar uma predição, e o resultado da predição do algoritmo é realizado por votação, ou *agregação*: uma combinação das predições de todas as árvores da floresta. Ao processo de *bootstrapping*, junto do processo de *agregação*, chama-se *bagging*.

A regressão logística, do inglês *logistic regression* (LR), é um algoritmo que não tem por origem a ciência da computação, e sim a estatística. Originalmente, a LR foi desenvolvida para preencher a lacuna de que todos os modelos aplicados de regressão linear produziam resultados insatisfatórios. A LR permite tanto a utilização de preditores contínuos quanto categóricos, fornecendo uma grande capacidade de ajuste para múltiplos preditores. Estes pontos tornam a LR muito útil para análise de dados observacionais – quando ajustes são necessários para reduzir potenciais diferenças nas classes que estão sendo comparadas.

Já as máquinas de vetores de suporte, ou *Support Vector Machines* (SVM), é um algoritmo que consegue traçar hiperplanos que servem como limite para as tomadas de decisão. Em um problema de classificação, por exemplo, quando existem instâncias pertencentes a diferentes classes, o maior hiperplano possível serve justamente para separá-las [67]. Caso essas instâncias não sejam linearmente separáveis, podem ser utilizadas funções matemáticas chamadas um pouco mais complexas, *kernels*, que elevam os dados em algumas dimensões até que um hiperplano consiga separar adequadamente as classes. Esta é uma grande vantagem deste algoritmo e o manteve como estado-da-arte durante praticamente uma década, antes do surgimento das redes neurais profundas. Além disso, pode-se citar como vantagens adicionais a capacidade de lidar com a manipulação de dados estruturados ou semiestruturados, além de resolver tanto problemas de classificação quanto de regressão.

Como uma generalização é adotada no SVM, há menos probabilidade de *overfitting*. Como desvantagens da utilização deste algoritmo, pode-se citar o baixo desempenho quando o volume de dados presente no conjunto utilizado é muito grande (ou quando há ruídos no conjunto de dados); o não-fornecimento de estimativas de probabilidade no resultado final; e a dificuldade de compreensão do modelo gerado ao final.

O algoritmo *Gaussian Naive Bayes* (NB) possui uma abordagem probabilística, derivado do Teorema de Bayes [21]. Para cada atributo presente no conjunto de dados, este algoritmo calcula a sua média e o seu desvio padrão, resultando na curva que representa a distribuição gaussiana destes dados – daí seu nome. Então, o passo seguinte é criação de uma tabela de probabilidades para cada classe, chamada de probabilidades *a priori*. Para as novas instâncias, este algoritmo multiplica a probabilidade *a priori* de cada classe

muito bem ao conjunto de dados que foi levado em consideração no treinamento e na validação, superestimando o desempenho do modelo gerado. Entretanto, se mostra ineficaz para prever novas observações.

com as probabilidades condicionais (*likelihood*³) de cada atributo. Matematicamente, em um plano cartesiano, *likelihood* pode ser entendida simplesmente como a coordenada y da curva que corresponde à coordenada x). Entretanto, para lidar com o problema de *underflow*⁴ ocasionado pela multiplicação de probabilidades, o algoritmo realiza somatório do logaritmo natural (base e) das probabilidades condicionais. A classificação da nova instância, através deste algoritmo, se dá pela classe com maior resultado.

Como exemplos clássicos de utilização deste algoritmo, pode-se citar os filtros de *spam* [20], no recebimento de e-mails; a mineração de emoções e a computação afetiva (identificação e detecção das emoções em textos) [68]; na separação de documentos, notícias [94]; etc.

O algoritmo *Adaptive Boosting*, conhecido popularmente pelo acrônimo *Adaboost* (AB), é um meta-algoritmo que combina diferentes classificadores fracos associados a diferentes pesos na tomada de decisão do algoritmo. Diferentemente do *Random Forest*, em que cada árvore criada na floresta não possui um tamanho pré-fixado, as árvores do *Adaboost* normalmente possuem pouquíssimos nós. O termo usual, em inglês, para este tipo de árvore que possui poucos nós é *stump*. Tecnicamente, estes *stumps* são chamados de *weak learners*.

No contexto do *Random Forest*, no momento da votação que define o resultado da classificação do algoritmo, o peso de cada árvore é exatamente igual. Já no *Adaboost*, os *weak learners* possuem pesos diferentes na tomada de decisão. Por último, no *Random Forest*, cada árvore da floresta pode tomar decisões de forma independente – as árvores de uma mesma floresta não se relacionam entre si. Nesse sentido, pode-se afirmar que o *Random Forest* não leva em consideração a ordem de criação das árvores da floresta, pois elas não se inter-relacionam de forma direta antes da votação final. No *Adaboost*, a ordem de criação dos modelos tem extrema relevância, pois cada árvore é criada de acordo com os erros normalizados das árvores anteriores, isto é, os erros da primeira árvore influenciam em como a segunda árvore aprende um modelo que intensifica o aprendizado das instâncias mais difíceis de determinar a classe.

O algoritmo *Gradient Boosting* (GB) também é um algoritmo do tipo *ensemble*, ou seja, um algoritmo que não possui uma classificação determinada, justamente por fazer parte da família de algoritmos de método combinado, de duas ou mais famílias distintas de algoritmos, a exemplo dos supracitados *Random Forest* e *Adaboost*. É um algoritmo que se caracteriza pela sua flexibilidade: além de resolver problemas de regressão, é capaz de lidar muito bem com problemas também de classificação.

O *Gradient Boosting* normalmente está inserido no contexto das árvores de decisão,

³*Likelihood* é um termo em inglês que, no português, possui o mesmo significado de *verossimilhança*. Entretanto, enquanto probabilidade significa a possibilidade de algo acontecer, *likelihood* faz referência ao processo de determinar a melhor distribuição possível dos dados, dada uma situação específica.

⁴Quando ocorrem multiplicações de números muito pequenos, o resultado leva a um número ainda menor, até que seja arredondado para zero.

e sua terminologia deriva de funções distintas – mas interpoladas – do próprio algoritmo: enquanto *boosting* vem da combinação de diversos e diferentes *weak learners*, conforme explicado anteriormente, o termo *Gradient* está presente pelo fato de este algoritmo basear-se nos resíduos⁵ e, mais precisamente, na derivada da função de perda.

Uma função de perda, ou *loss function*, tem sempre duas entradas: a classe verdadeira e a classe predita – esta última sendo a saída do modelo desenvolvido. De forma simplificada, quanto maior a função de perda, pior é o desempenho do modelo de predição.

Inicialmente, este algoritmo cria um *weak learner* que, naturalmente, ainda está muito longe de poder prever, com maior segurança, classes verdadeiramente positivas. Então, ele verifica quais as chances matemáticas de que cada instância tem de ser classificada como uma determinada classe. Pode-se expressar as chances iniciais de cada instância da seguinte forma:

$$PI = \frac{1}{|classes|}$$

onde PI é o valor da predição inicial (chances) e $|classes|$ é o número de classes únicas e distintas possíveis de terem instâncias classificadas. Então a primeira folha da árvore, representando a predição inicial, é igual ao logaritmo das chances iniciais de uma instância ser de uma determinada classe.

Após a suposição da predição inicial, para cumprir o objetivo da classificação, o algoritmo converte o resultado do logaritmo em uma probabilidade, e calcula o resíduo entre esta probabilidade e o valor predito. O passo seguinte realizado pelo algoritmo é o cálculo

$$\frac{\sum res_{folha}}{\sum prob_{apriori} \times (1 - prob_{apriori})}$$

onde res_{folha} são os resíduos de cada folha e $prob_{apriori}$ são as probabilidades *a priori* preditas da folha (o resultado do logaritmo, a suposição inicial do algoritmo). O resultado da fórmula acima funciona, para este algoritmo, como uma espécie de “peso” para cada folha do nó desta árvore, e será utilizado combinando a probabilidade *a priori* (o resultado do logaritmo) com a nova árvore. Entretanto, a grande diferença se dá pelo fato de o *Gradient Boosting* utilizar, ainda, uma *learning rate*, ou *taxa de aprendizado*, que irá modificar a escala dos valores de saída de cada folha da árvore, fornecendo, ao novo modelo, uma certa “quantidade” do novo *weak learner* que será adicionada ao *weak learner* anterior. Daí será originado um novo *weak learner*, com um desempenho potencialmente maior que o anterior.

Este processo se repete tantas vezes forem necessárias, até que se atinja o número de árvores – definido como hiperparâmetro – pelo usuário, ou até que o valor residual

⁵Tomando por exemplo uma equação de regressão, os resíduos seriam as distâncias entre os pontos reais e os pontos preditos pela função.

seja muito pequeno. A cada *weak learner* adicionado ao *weak learner* anterior, o modelo dá um pequeno passo na direção em que a perda é minimizada, justamente através da sua derivada – ou gradiente. Ao final, se o valor de probabilidade (o qual chama-se probabilidade *a posteriori*) for superior ao limite pré-estabelecido (*threshold*), a instância é classificada como uma verdadeira positiva.

Embora o *Gradient Boosting* tenha um desempenho muito bom nas mais variadas classes de tarefas, uma das desvantagens da utilização deste algoritmo é em relação à escalabilidade, principalmente em função do custo computacional.

Por último – mas não menos importante – é preciso explicar sobre a classe de algoritmos que representam as *Artificial Neural Networks* (ANN), ou *Redes Neurais Artificiais*, mais especificamente o *Multilayer Perceptron* (MLP).

As Redes Neurais, um dos mais populares algoritmos em se tratando de *Machine Learning*, representam uma série de conceitos e técnicas distintas. São os modelos popularmente conhecidos como *caixa preta*, pois seu modelo final pode ser difícil de ser interpretado.

A base biológica das redes neurais tem, por pano de fundo, uma controversa⁶ tentativa de modelar o funcionamento do cérebro humano. O cérebro humano consiste, basicamente, em bilhões de células individuais, chamadas de *neurônios*. Tem-se por postulado que todo conhecimento e experiência adquiridos durante a vida humana são codificados através das conexões existentes entre os neurônios. Na grande área da biologia, tais conexões são chamadas *sinapses*.

Como a discussão sobre a controversa ideia que motivou a criação das redes neurais não é o objetivo deste trabalho, de forma simplificada – e cabendo perfeitamente uma comparação com o funcionamento biológico do cérebro humano – pode-se afirmar que uma rede neural artificial é composta de nós e conexões entre os nós, sendo os nós iniciais da rede neural compõem as entradas do modelo, enquanto que os nós finais, a saída do modelo. Aos nós, chama-se *perceptron*. As conexões entre os *perceptrons* são associadas a determinados pesos. A saída de uma rede neural é composta das previsões que o modelo é capaz de realizar.

A estrutura básica de uma rede neural, então, consiste em *perceptrons* (que podem ser comparados, remontando à ideia biológica, a neurônios artificiais) e pesos. Cada *perceptron* realiza determinadas operações sobre os dados que chegam até ele – as entradas recebidas através das suas conexões anteriores. O comportamento resultante de uma rede neural artificial é decorrente, então, das interações entre os vários *perceptrons*.

⁶Alguns autores entendem que todo este *background* biológico atribuído às redes neurais artificiais foi enviesado no funcionamento deste modelo de forma intencionalmente comercial. Rauber [66] entende que os modelos de redes neurais artificiais têm pouco em comum com as redes neurais reais. Pulvermüller et al. [62] entende que a similaridade das redes artificiais aos aspectos relativos à anatomia e à fisiologia do cérebro são imperfeitas, enquanto Mazzoni et al. [51] entende que o aprendizado por retropropagação, o método geralmente usado para treinar redes neurais artificiais, é claramente não-biológico.

A arquitetura típica de uma rede neural artificial é composta de três diferentes camadas de *perceptrons*: a *input layer* (camada de entrada), a *output layer* (camada de saída) e, entre elas, a *hidden layer* (camada oculta). A camada de entrada é responsável por receber o sinal de entrada, ou seja, os dados que irão ser processados pela rede neural. A tarefa que a rede neural se destina – a predição – é executada pela camada de saída. As camadas ocultas, que consistem em um número arbitrário que pode ser definido pelo próprio desenvolvedor do modelo, colocadas entre as camadas de entrada e saída, são o verdadeiro mecanismo computacional de uma *Multilayer Perceptron* [59]. Pode-se afirmar, portanto, que o processo de aprendizado de uma rede neural artificial reduz-se a encontrar os valores corretos para os pesos entre a camada de entrada e a camada de saída.

Ao processo de encontrar os melhores valores para os pesos, dá-se o nome de *learning rule*, ou *regra de aprendizado*. Tal regra é estabelecida sobre o ponto nevrálgico das redes neurais – o algoritmo de retro-propagação, ou *backpropagation*. Este algoritmo é composto de duas fases distintas: propagação e retro-propagação.

Inicialmente, os pesos estabelecidos em uma rede neural são estabelecidos de forma aleatória e, nesse momento, não se tem certeza de que os seus valores serão suficientes para atender com satisfatoriedade um determinado problema. Então, o processo de propagação é realizado. Ao final, o erro é verificado. Em via de regra, os valores iniciais dos pesos não são suficientes, e um processo de ajuste dos valores estabelecidos aleatoriamente precisa ser realizado. Este ajuste dos valores dos pesos é realizado pela segunda fase do algoritmo, retro-propagando o erro através da rede neural, partindo da camada de saída, passando novamente pelas camadas ocultas, até a camada de entrada. Este ciclo é repetido tantas vezes forem suficientes – até que o erro encontrado pela rede neural seja o menor possível.

Então, pode-se afirmar que o *aprendizado* de uma rede neural se resume ao processo de encontrar os melhores valores para os pesos, até que seja decidido por uma *função de ativação* que a rede resolveu o problema.

2.5 Lidando com o desbalanceamento de classes

Em situações cotidianas, observadas do mundo real, do dia-a-dia, o desbalanceamento de classes é tido como uma característica natural e inerente dos dados [39]. Dados desbalanceados representam um ponto de inflexão e merecem uma grande atenção no momento do desenvolvimento dos modelos de classificação [38], pois o processo de treinamento dos próprios modelos poderá ser totalmente enviesado, podendo, até, em determinadas situações, fazer com que o modelo final ignore por completo a classe minoritária.

Oversampling é uma técnica de amostragem usada para corrigir o desbalanceamento de classes que visa aumentar o número de instâncias das classes minoritárias de forma artificial. Reduzir a quantidade de instâncias da classe majoritária é chamado de *under-*

sampling. Quando as instâncias são escolhidas aleatoriamente sem a criação de dados sintéticos diferentes das instâncias originais, os métodos são também conhecidos como *upsampling* e *downsampling*, respectivamente.

Quando apenas a réplica aleatória de instâncias reais da classe minoritária não altera a capacidade de generalização dos modelos de forma satisfatória, podem ser adotados outros métodos mais complexos, tal como o algoritmo *Synthetic Minority Over-sampling Technique* (SMOTE) [14].

O SMOTE foi desenvolvido para lidar, exclusivamente, com o problema de desbalanceamento de classes em conjuntos de dados, criando instâncias sintéticas das classes minoritárias. O funcionamento geral deste algoritmo pode ser descrito, de forma intuitiva, seguindo alguns passos, resumidos a seguir: (i) escolher aleatoriamente uma instância da classe minoritária; (ii) selecionar os k-vizinhos mais próximos da instância escolhida; (iii) escolher aleatoriamente um dos k-vizinhos selecionados; (iv) gerar a instância sintética da classe minoritária, combinando a instância e o k-vizinho selecionados. Assim, as instâncias da classe minoritária não são, simplesmente, replicadas. São criadas novas instâncias. Como vantagem principal do SMOTE, destaca-se a minimização da ocorrência de *overfitting* causado pelo *oversampling*. Por este motivo, o algoritmo deve ser considerado quando há muito menos instâncias nas classes minoritárias, como é o caso deste trabalho. Entretanto, tem como desvantagem a possível introdução de ruído nos dados.

3 TRABALHOS RELACIONADOS

Para garantir o desenvolvimento de uma metodologia sólida de predição e, desta forma, sugerir uma solução para o problema proposto, fez-se necessário um levantamento bibliográfico, ainda que de forma não-sistemática, dos trabalhos relacionados com o tema do presente estudo inseridos na literatura científica relevante, buscando evidenciar diferentes métodos, relacionados de forma direta e indireta com a resolução do tema de pesquisa em questão. Sendo assim, esta seção tem por objetivo apresentar os trabalhos relacionados com o presente objeto de pesquisa. Para melhor estruturação e entendimento, dividimos esta seção em três categorias: modelos baseados em Aprendizado de Máquina (Seção 3.1); modelos analíticos, estatísticos, ou baseados em generalizações (Seção 3.2); e ainda outros trabalhos correlatos (Seção 3.3).

3.1 Modelos baseados em Aprendizado de Máquina

A cadeia de suprimentos de produtos florestais, na qual a exploração da madeira em toras está inserida, dá origem a uma variedade de problemas interligados. Nesse sentido, abordagens com algoritmos de *Machine Learning* têm sido utilizadas para prever o teor de umidade da biomassa de madeira como um parâmetro crítico para controlar as qualidades e propriedades de biocombustíveis.

Quando uma antena é fixada dentro de um montante de cavacos de madeira, sua impedância de entrada varia de acordo com a permissividade da madeira, permitindo medir o coeficiente de reflexão. Então, um método baseado em medição de radiofrequência foi obtido [16]. Assim, as saídas do modelo do coeficiente de reflexão são usadas como entradas no modelo de teor de umidade.

O modelo, que foi então gerado, permite prever o teor de umidade da madeira a partir de valores estimados do coeficiente de reflexão utilizando métodos de seleção de características e Máquinas de Vetores de Suporte baseadas no método dos mínimos quadrados (*Least Squares Support Vector Machines* – LS-SVM) [81]. Duas espécies de madeira foram envolvidas: *pinus* e *carvalho*. Uma pilha de cada espécie estava disponível neste experimento. Num primeiro momento, as pilhas foram umedecidas e, em seguida, a

mudança no teor de umidade foi monitorada ao longo do tempo.

Os autores do estudo testaram dezesseis diferentes valores de teor de umidade. Os resultados demonstram a eficácia da metodologia de modelagem proposta no referido estudo, permitindo que a solução para predição do teor de umidade seja adequada para implantação direta em processos industriais que requeiram madeira para produção de energia em tempo real.

Outra abordagem usando aprendizado de máquina, mais especificamente um método de modelagem de teor de umidade, foi proposta usando o algoritmo *Support Vector Machines* (SVM) [91]. Para este modelo, os atributos de entrada foram a temperatura e o teor de umidade de equilíbrio, ou seja, o teor de umidade no qual a madeira não ganha nem perde umidade. A saída do modelo é o próprio teor de umidade. Os dados de treinamento foram obtidos em um forno de secagem de madeira com temperatura entre 26,6 e 37,7 °C e aquecimento a vapor. Além de utilizar o forno de secagem, ainda havia equipamentos de aquecimento e instrumentos de medição, incluindo sensores de temperatura, sensores de umidade, sensores de umidade de equilíbrio, ventiladores de seca e controladores. Foi realizada uma seleção de parâmetros experimentais para o modelo. A técnica de filtro *infinite impulse response* (IIR) foi aplicada sobre o ruído gerado nos dados de treinamento. Os resultados deste estudo validaram a eficácia do método que foi proposto, demonstrando que é possível utilizar aprendizado de máquina para esta finalidade.

A literatura científica também aborda aplicação de lógica *fuzzy* para prever o teor de umidade em madeira [7]. Neste trabalho, o teor de umidade foi predito em função da temperatura e do tempo de secagem. Os autores utilizaram a espécie *Fagus orientalis*, que é uma madeira comumente utilizada em processos fabris, em secagem natural. As dimensões das amostras foram de 55×25×25mm. Em seguida, as amostras foram mantidas na sala climatizada até que a variação de peso atingisse o equilíbrio. Após este processo, o teor de umidade de 13% foi alcançado. Em seguida, as amostras de madeira foram secas em uma estufa de secagem em diferentes temperaturas (50, 70, 90 e 110°C) e tempo (0,5, 1,0, 1,5, 2,0, e 2,5 horas). Dez amostras foram preparadas para cada experimento. Após ajustado, o modelo apresentou uma acurácia média de 97,16%, demonstrando que a lógica *fuzzy* pode ser utilizada como uma ferramenta valiosa no processo de predição de secagem da madeira – uma importante parcela do custo da indústria madeireira.

Algoritmos de *Machine Learning* também foram utilizados para se obter a predição de volumes de madeira em árvores e melhor alocação de toras de madeira para diferentes usos, comparando-os com as equações de afilamento da madeira [77]. As equações de afilamento são expressões matemáticas que explicam a relação de diâmetro com a altura de uma árvore, ou de uma população de árvores. Diferentemente das abordagens tradicionais, em que modelos de cálculo de volume de madeira em árvores eram baseados apenas em uma única medida do diâmetro da árvore, as equações de afilamento – através das medidas de conicidade dos troncos das árvores – podem retornar estimativas de vo-

lume muito mais precisas, levando em consideração no cálculo as mudanças de diâmetros decrescentes da base até o topo.

No estudo supracitado, comparou-se diferentes algoritmos para o problema, como Redes Neurais Artificiais (*Artificial Neural Networks* – ANN), Regressão de Vetor de Suporte (*Support Vector Regression* – SVR) e Floresta Aleatória (*Random Forest* – RF), buscando entender o comportamento das previsões dos algoritmos, além de especificar os melhores modelos para os casos estudados. Os autores demonstraram que a abordagem utilizando ANN apresentou as estimativas com maior acurácia. Já a abordagem utilizando RF gerou estimativas imprecisas, a não ser que as mensurações sejam tomadas em intervalos menores e com grandes amplitudes de classes de tamanho de toras. Ao final, conclui-se deste estudo que os algoritmos testados mostraram um desempenho igual ou superior ao das abordagens convencionais.

Uma das finalidades da madeira armazenada em toras é ser matéria-prima em processos industriais a partir de serrarias. O objetivo das serrarias é transformar a madeira em toras (matéria-prima) em madeira serrada ou laminada (produto intermediário ou, dependendo do caso, produto acabado).

Para melhor organizar as toras de madeira utilizadas como matéria-prima nas serrarias, pesquisadores demonstram como modelos de aprendizado de máquina podem fornecer recomendações para várias tarefas, tanto na alocação de blocos de corte, auxílio na etapa de planejamento e recomendação tanto da madeira quanto da fábrica responsável pelo processamento [56]. O estudo permitiu aprimorar a produção das indústrias reduzindo perdas dos processos. Os modelos propostos alcançaram resultados de até 94% do ganho teórico máximo em um contexto de tomada de decisão. No caso estudado, é impossível superar o desempenho dos modelos do simulador, uma vez que estes modelos foram utilizados na etapa de treinamento do próprio modelo. Ainda assim, a abordagem de aprendizado de máquina que leva ao maior ganho são as árvores de decisão (*Decision Trees* – DT), com um ganho médio de 83% em relação ao método mais usual da produção histórica.

Com relação à variação do teor de umidade, pode-se citar um trabalho que também se utilizou de Redes Neurais Artificiais para prever as taxas de secagem de fornos para madeira, com base nas espécies e em informações básicas de densidade da madeira [93]. Ainda que esta não seja exatamente uma sub-área de interesse da presente pesquisa – por se tratar de secagem em estufa, e não de secagem natural – a contribuição para a estimativa do teor de umidade da madeira utilizando algoritmos de aprendizado de máquina deve ser considerada.

A rede neural desenvolvida contava com três entradas: teor de umidade inicial, densidade e tempo de secagem. A saída do modelo foi a estimativa do teor de umidade final médio. Um algoritmo de retro-propagação foi implementado para treinar, validar e testar o modelo com 50, 25 e 25% de instâncias, respectivamente.

Após obter a configuração otimizada variando os principais hiperparâmetros, como a função de transferência, a taxa de aprendizado e o número de neurônios e camadas de aprendizado, a qualidade do modelo final atingiu um Erro Relativo Absoluto Médio (MRAE) inferior a 2%, confirmando a excelente capacidade preditiva deste método para prever a taxa de secagem. Sabe-se que uma grande vantagem da utilização de Redes Neurais Artificiais em comparação com modelos mais triviais de regressão, por exemplo, é justamente a liberdade quanto às suposições lineares, sendo potencialmente mais eficazes para lidar com não-linearidades. Em contrapartida, a baixa explicabilidade dos modelos finais gerados é um ponto negativo a ser igualmente considerado.

3.2 Modelos analíticos, estatísticos, ou baseados em generalizações

Modelos baseados em métodos analíticos ou estatísticos já foram utilizados para desenvolver, comparar e validar diferentes abordagens sobre o problema de interesse deste trabalho de pesquisa aplicada.

Na Finlândia, pesquisadores estimaram o tempo ideal de armazenamento de pilhas de madeira mantidas ao ar livre com base nas mudanças de teor de umidade [65]. Modelos multivariados foram criados com o objetivo de estimar as mudanças no teor de umidade em diferentes ambientes de secagem, com base em modelos de teste e validação desenvolvidos para as espécies de madeira *pinus*. Esses modelos foram aplicados a pilhas da espécie *pinus silvestre*, que são destinadas à produção de combustível. Para atender diferentes níveis de umidade ao longo dos diferentes períodos do ano, os dados do experimento foram coletados durante um período de sete a quatorze meses e consideraram a precipitação, a evaporação e a espécie da madeira como fatores principais.

O estudo comparou modelos de regressão, para estimar o teor de umidade, com o método mais tradicional – a diferença periódica de peso das toras armazenadas em pilhas. A conclusão foi que o registro de dados baseado em células de carga para estimar a mudança no teor de umidade da madeira mostrou-se uma opção viável para obter os dados necessários para os modelos de secagem da madeira, ou seja, a proposição deste estudo para avaliar a diferença do teor de umidade da madeira em um determinado período de tempo foi a pesagem periódica do produto em questão.

Com relação ao processo de secagem da madeira, existem pesquisas diretamente relacionadas à construção de métricas para diferentes estimativas. O tema vem sendo pesquisado desde 1959, e sabe-se que é regido por algumas variáveis, como temperatura, umidade relativa, velocidade do vento e precipitação [44]. Em 2009, outro estudo inseriu mais variáveis a serem estudadas e pesquisadas, como acesso a correntes de vento e exposição ao sol [43] no processo de secagem ao ar livre, também conhecido como *secagem natural*.

Em 1982, pesquisadores promoveram um calendário de secagem natural, baseando-se

em dados meteorológicos e na perda de umidade, para duas espécies de madeira: *carvalho vermelho* e *choupo amarelo*, ambas com 25mm de espessura [18]. As amostras foram expostas às condições ambientais em três diferentes pátios de armazenamento, por um período de cinco meses. O modelo desenvolvido pelos autores do estudo foi capaz de prever a perda diária do teor de umidade e incluiu outras variáveis até então inexploradas, como temperatura, umidade relativa média diária e teor de umidade médio diário. Outras variáveis meteorológicas também foram utilizadas, como velocidade do vento e precipitação, mas concluiu-se que não eram significativas. A taxa diária de perda de umidade foi relacionada a dados obtidos de uma estação meteorológica regional. Como conclusão, o estudo mostra que o modelo de regressão que melhor previu a perda diária estimada de teor de umidade foi:

$$\delta_M = a + bM^n + cT + dH,$$

onde δ_M é a variação do teor de umidade; M é o teor de umidade no início do dia; T é a temperatura média diária, medida em graus Fahrenheit; H é a umidade relativa média diária; a , b , c e d são coeficientes de regressão; $n = 1$ para *choupo amarelo* e $n = 2$ para *carvalho vermelho*. Este resultado específico permitiu estimar o tempo de secagem natural para qualquer data de empilhamento durante o ano, bastando que os dados locais de temperatura e umidade relativa estivessem disponíveis.

Em 2000, pesquisadores utilizaram uma abordagem distinta, uma simulação computacional, para desenvolver uma forma analítica de estimar os tempos de secagem natural de várias espécies de madeira laminada a partir de dados meteorológicos locais para madeira empilhada em qualquer dia do ano [73]. Neste estudo, também foi incluído o efeito da espessura da madeira laminada no tempo de secagem. O modelo apresentado se baseia em tempos experimentais de secagem natural para seis espécies de madeira: *carvalho vermelho do norte*, *bordo de açúcar*, *faia americana*, *álamo amarelo*, *pinus ponderosa* e *abeto de Douglas*.

A premissa deste estudo foi que, uma vez encontrados os parâmetros para cada espécie de madeira na localização geográfica dos dados experimentais, eles pudessem ser utilizados na simulação de secagem, estimando os tempos de secagem natural em outros locais onde os dados meteorológicos históricos também estivessem disponíveis. Os resultados mostram os tempos estimados de secagem natural das seis espécies de madeira, para qualquer espessura, até qualquer teor de umidade final, em qualquer dia do ano e em qualquer local onde existam dados de temperatura média e umidade relativa.

Em 2003, os valores do modelo de regressão proposto no estudo datado de 1982 foram ajustados, mas agora para estimar os tempos de secagem natural em toras com diâmetros entre 10,2 cm e 20,3 cm [75]. Assim, foi possível desenvolver modelos que relacionavam a perda diária do teor de umidade com o teor de umidade no início do dia, com a tem-

peratura média diária, umidade relativa e diâmetro da tora de madeira. Para este estudo, é essencial mencionar que, embora as toras de madeira tenham sido expostas à secagem natural, elas não foram expostas às intempéries naturais, pois foram revestidas com uma placa de compensado, protegendo-as tanto da chuva quanto da exposição direta ao sol. Os dados meteorológicos para este estudo foram monitorados e registrados usando um registrador de dados alimentado por bateria montado próximo às pilhas, medindo temperatura e umidade relativa a cada dez minutos e armazenando os dados junto com a hora e a data. Após o período de estudo, os dados de temperatura e umidade relativa foram então calculados para cada dia para serem usados na posterior análise dos dados.

A partir deste trabalho, que introduz um modelo de regressão para toras de madeira, duas equações diferentes foram analisadas:

$$\delta_M = a + bM + cT + dH + eD,$$

$$\delta_M = aM^bT^cH^dD^e,$$

onde δ_M é o percentual de perda diária do teor de umidade; M é a porcentagem do teor de umidade no início de qualquer dia durante o processo de secagem; T é a temperatura média diária, medida em Fahrenheit; H é o percentual de umidade relativa média diária; D é o diâmetro da tora em polegadas; a , b , c , d e e são coeficientes determinados por regressão. Os dados coletados nos experimentos foram ajustados por regressão linear múltipla (primeira equação) e por regressão não-linear (segunda equação). Para definir o melhor modelo, foi utilizado o coeficiente R^2 . Como resultado, o modelo não-linear obteve melhor ajuste para a espécie *pinus ponderosa*, com $R^2 = 0,786$, e o modelo linear para a espécie *abeto de Douglas*, com $R^2 = 0,723$.

Em 2004, um estudo [74] do mesmo pesquisador desenvolveu um modelo de regressão não-linear para descrever a secagem natural igualmente para duas espécies de madeira. Nesta pesquisa, foram desenvolvidos coeficientes de regressão para prever a perda diária do teor de umidade com base nas mesmas variáveis: teor de umidade no início do dia, temperatura e umidade relativa média diária. As espécies de madeira utilizadas neste estudo foram *carvalho vermelho do norte* e *ácer açucarado*. Os modelos de regressão resultantes são apresentados abaixo:

$$\delta_M = MC^{2.38}F^{0.759}RH^{-2.91}(\text{carvalho vermelho do norte})$$

$$\delta_M = MC^{2.18}F^{1.89}RH^{-3.57}(\text{ácer açucarado})$$

Os modelos podem prever a perda diária do teor de umidade com base em praticamente as mesmas variáveis (MC é o teor de umidade inicial da madeira, F é a temperatura média diária e RH é a umidade relativa média diária). Portanto, tendo as variáveis em comum, as equações poderiam ser utilizadas para desenvolver curvas de perda do teor

de umidade ao longo do tempo, bastando que os dados meteorológicos locais estivessem disponíveis, observando-se apenas as espécies de madeira.

A pesquisa supracitada resultou no desenvolvimento de uma planilha eletrônica para estimar o tempo necessário para a secagem natural da madeira, com base em um índice denominado *Air-Drying Estimator* [53]. O modelo requer os seguintes dados de entrada: tipo de madeira (*carvalho vermelho*, *álamo amarelo* ou *ácer açucarado*); a localização do pátio; a data de início da secagem da madeira; o teor de umidade inicial; e o teor de umidade-alvo que se deseja atingir. A saída do modelo contém: a data estimada em que a madeira atingirá o teor de umidade-alvo desejado pelo usuário; o número de dias necessários para atingir o teor de umidade-alvo; o número médio de dias de secagem, assumindo que a secagem começou no primeiro dia de cada mês ao longo do ano; a curva de secagem exibindo o teor de umidade e os dias de secagem; a comparação das curvas de secagem para cada mês; o histograma comparando o número de dias de secagem necessários.

Este modelo evidencia a influência significativa do teor de umidade inicial para a estimativa final da perda do teor de umidade diária. Isso ocorre porque a taxa de secagem inicial da madeira laminada verde é muito maior do que a taxa durante os estágios posteriores de secagem.

Em seguida, o modelo criado foi aplicado a outros modelos já desenvolvidos anteriormente, especialmente o estudo de 1982. As equações do modelo foram desenvolvidas com base em dados diários do clima e teor de umidade da madeira, mas o modelo utiliza dados meteorológicos médios mensais para calcular a perda diária de umidade. Apesar do esforço e da notável evolução trazida por este modelo, é permitido ao usuário empregar diferentes equações para localizações geográficas distantes de onde foram desenvolvidas, o que pode induzir o próprio modelo a erros significativos. Em relação às espécies de madeira, os agrupamentos também têm por base outros trabalhos publicados.

Corroborando com os estudos acima, outro estudo demonstra que dados meteorológicos utilizados em modelos de secagem natural podem estimar adequadamente o teor de umidade da madeira em toras [23]. Além disso, considerando que determinar o teor de umidade em uma pilha de toras de madeira, sem medi-la com frequência, é um problema operacional, este estudo teve como objetivo modelar a mudança no teor de umidade de uma pilha de toras de madeira em função de variáveis de entrada mensuráveis. Assim, os períodos de secagem natural puderam ser otimizados e, assim, pudesse ser determinado o período de transformação da madeira em lenha (para produção de energia).

Para estimar esses dados, utilizou-se um equipamento para medir constantemente – a cada dez minutos – as amostras de madeira em toras. Os dados coletados foram: percentual de umidade relativa, temperatura, velocidade do vento e precipitação. Para efeitos de cálculo, mesmo que fosse na forma de neve – precipitação não líquida, a precipitação foi transformada no equivalente à precipitação líquida. O modelo de regressão gerado foi:

$$CMC = 4,691 \times 10^{-3}DP + 1,359 \times 10^{-2}P,$$

onde CMC é o percentual diário de variação do teor de umidade, T é a temperatura (medida em graus Celsius), DP é o percentual de umidade relativa diária potencial e P é a soma, em milímetros, da precipitação líquida diária e não-líquida.

O objetivo deste modelo de regressão múltipla foi basear-se em médias e somas diárias, buscando operacionalizar modelos previamente estudados de forma mais fácil e rápida, não necessitando de dados das toras em si. Uma importante informação levantada por este estudo diz respeito às espécies de madeira. Os resultados dos dados de precipitação não coincidem com outros estudos anteriores, que afirmavam que a precipitação não teve efeito significativo na secagem natural [31]. Assim, assume-se que certas espécies de madeira, como *pinus selvagem* e *salgueiro*, possuem apenas uma umidade superficial. Além disso, a espécie *pinus selvagem* pode ser considerada mais propensa a reter umidade na casca do que a espécie *salgueiro*.

3.3 Outros trabalhos correlatos

Sobre o armazenamento de toras de madeira, um estudo investigou o efeito causado pelo tempo de armazenamento de toras de madeira na Nova Zelândia [89]. Neste trabalho foram realizados dois ensaios representando condições favoráveis e desfavoráveis de armazenamento: no verão, em local quente e seco; e no inverno, em local frio e relativamente úmido. Foram instaladas 20 pilhas, contendo aproximadamente 600 kg de toras de madeira cada (peso inicial úmido). O teor de umidade foi determinado gravimetricamente no início e no final dos testes. Todas as pilhas foram pesadas em intervalos entre uma e quatro semanas para identificar tendências de perda de peso ao longo do tempo de armazenamento.

Após vinte e quatro semanas de armazenamento no verão, o teor de umidade das toras de madeira diminuiu de um valor inicial de 53% para valores entre 33 e 21%. O decréscimo foi mais significativo para as toras pequenas e descobertas, e menor para as toras grandes e cobertas. Devido às condições climáticas úmidas e frias, as toras armazenadas no inverno secaram muito pouco em dezessete semanas. O teor de umidade diminuiu de um valor inicial de 58% para valores entre 51 e 49%, sem diferenças significativas de tratamento observadas no teste de inverno.

Concluiu-se que a técnica de armazenamento de toras de madeira através do empilhamento de pequenas toras, sem qualquer cobertura, é a técnica que mais otimiza o processo de secagem da madeira. As toras maiores secavam mais lentamente, e a divisão das pilhas acelerou sua secagem. Concluiu-se, também, que a cobertura das pilhas não ajuda a

diminuir a perda de umidade, e os resultados também indicaram que a cobertura das toras não acelerou a secagem das toras molhadas.

A madeira pode ser armazenada em pilhas para diversas finalidades. Uma delas é para a produção de energia. Um estudo produzido na Suécia [84] analisou a relação entre a perda de umidade e a perda na produção de energia. Para medir e determinar as mudanças na produção de energia em razão do armazenamento, foram analisadas a umidade e a perda de matéria (perda de peso) antes e após o armazenamento. Este estudo concluiu que se a madeira armazenada em pilhas for reduzida a cavacos ao invés de toras de madeira, a perda de energia pode variar entre 7% e 21% por um período entre seis e nove meses de armazenamento.

Nesse mesmo sentido, e com o mesmo objetivo, outro estudo apresenta resultados semelhantes, com maior detalhamento em função do comprimento das toras armazenadas em pilhas [86]. O trabalho analisou as mudanças no teor de umidade e massa de madeira armazenada em toras por três meses. As toras de 1,2 m de comprimento foram armazenadas em pilhas, enquanto as toras de 9,6 a 14,4 m foram armazenadas no local da colheita, mas não foram empilhadas.

O estudo demonstrou que o teor de umidade das toras de 1,2 m armazenadas em pilhas diminuiu 22,9%, enquanto as toras armazenadas no local da colheita diminuíram 15%. Os resultados mostram que quanto maior o comprimento das toras armazenadas nas pilhas, menor a perda de umidade durante o período de armazenamento.

Sabe-se que a madeira de árvore que foi recentemente derrubada possui alto teor de água e que a perda de água é um fator preponderante para a perda do teor de umidade. Um estudo [85] teve como objetivo demonstrar o período mínimo para que as toras de madeira armazenadas em pilhas iniciassem o processo de secagem natural. Para determinar como o método e as condições de armazenamento afetam as mudanças de peso e a perda de umidade da madeira, foi selecionada uma parcela de sessenta árvores modelo, divididas em dois grupos: no primeiro grupo, trinta árvores inteiras; no segundo grupo, trinta árvores foram cortadas transversalmente, formando toras com 2,5 m de comprimento e armazenadas em pilhas.

A partir da experimentação neste estudo, notou-se que a madeira armazenada em pilhas perdeu umidade mais lentamente do que a madeira de árvores que foram deixadas inteiras após o corte. Comparando os pesos das toras armazenadas em uma pilha nos dias após a colheita, foi encontrada diferença estatisticamente significativa apenas entre o primeiro e o décimo quinto – e último – dia. Portanto, o estudo concluiu que o período de duas semanas é o período mínimo de armazenamento da pilha necessário para atingir um grau significativo de mudança de peso e perda de umidade.

Objetivando analisar o impacto da variabilidade nas propriedades da madeira na qualidade e no custo da celulose – outro produto derivado da utilização da madeira como matéria-prima, um estudo [72] foi realizado visando resolver o problema notado pela

indústria de celulose: os registros contábeis demonstravam que as toneladas de madeira comprada, divididas pelas toneladas de celulose produzidas, mudavam. O estudo, então, examinou o impacto do tempo – desde o momento em que uma árvore é derrubada na floresta até o momento em que ela é levada para a fábrica – sobre o rendimento e o custo operacional.

O percentual de perda de peso, assumido como mudança no teor de umidade, foi determinado para três diferentes espécies (a saber: *Pinus Taeda*, *Eucalyptus Dunni* e *Eucalyptus Grandis*) ao longo de doze semanas, tanto no inverno quanto no verão. Algumas amostras de madeira foram obtidas no momento imediatamente posterior à derrubada da árvore. Este percentual de perda de peso nos primeiros cinco dias de armazenamento variou de 2% a 13%, dependendo da espécie e da estação, e atingiu 36% após onze semanas de armazenamento. Em uma indústria que produz costumeiramente mil toneladas/dia, a diferença entre o uso de madeira de árvore recém cortada (verde) em função do uso de madeira cortada e armazenada por onze semanas é equivalente a 1.130 tons de madeira verde. Concluindo o estudo, presumindo que a indústria pague US\$ 45 por tonelada de madeira verde, a mudança no uso da tora de madeira verde e as toras com onze semanas de corte pode custar, a esta mesma indústria, aproximadamente US\$ 50.850 por dia.

Em se tratando da produção de madeira serrada e laminados de madeira, um estudo investigou o potencial de uso de diferentes espécies de eucalipto (*benthamii*, *deanei*, *E. dorrigoensis*, *E. dunni* e *E. smithii*) em uma região com ocorrência de geada [90]. As árvores selecionadas para amostragem tinham dezoito anos e foram seccionadas em toras de 2,1 m de comprimento. Em seguida, as amostras foram subdivididas em classes de diâmetro: entre 20,1 cm e 25 cm; entre 25,1 cm e 30 cm; e maiores que 30 cm. As rachaduras foram medidas, e então o rendimento da madeira pôde ser calculado.

Os resultados obtidos neste estudo indicaram que as espécies de madeira estudadas são de boa qualidade, destacando *E. dunni*. No entanto, as demais espécies também apresentaram rendimentos satisfatórios, indicando que podem ser utilizadas para a produção de madeira serrada em regiões com ocorrência de geadas. Para o presente levantamento bibliográfico, uma contribuição destacada desta pesquisa específica é a relação das rachaduras com o rendimento da madeira a ser laminada. Esse rendimento é expresso em um percentual obtido pela razão entre o volume de madeira laminada e o volume de toras expostas ao processo. Para o eucalipto, esse percentual varia entre 40 e 65%. Dentre alguns fatores que influenciam diretamente no rendimento da madeira, além das características da própria espécie, estão também a qualidade e o diâmetro das toras a serem processadas [87]. Por esse motivo, a divisão das toras em classes por diâmetro é uma importante estratégia para a produção de madeira laminada com alto rendimento.

3.4 Comparativo dos estudos relacionados no levantamento bibliográfico

Esta seção tem por objetivo comparar as diferentes abordagens analisadas durante o levantamento bibliográfico. Considerando que se forneceu o máximo possível de informações importantes por estudo (dezenove diferentes estudos), para facilitar a compreensão do resumo, dividiu-se a análise em duas partes, Tabelas 3 e 4. Em cada tabela, forneceu-se diferentes estudos – um por linha – e suas diferentes características por coluna. Destaca-se que a primeira coluna, ID, está relacionada à identificação do estudo considerado, portanto se repete em ambas as tabelas.

A Tabela 3 apresenta o ano da publicação, o tipo de madeira relacionada ao estudo, a espécie, o país e o objetivo de cada estudo relacionado ao teor de umidade. Além disso, considerando os estudos publicados, percebe-se que esse tema tem ganhado atenção mais recentemente, sendo um problema com um campo de pesquisa interessante. Também podemos observar que a América do Norte fornece a maioria dos estudos, sendo que os Estados Unidos possuem cerca de 30% das análises (6 casos), seguido pelo Canadá (3 casos). Destaca-se também os diferentes estudos que foram realizados na Europa. No que diz respeito às motivações, podemos observar que objetivos são otimização operacional, econômicos, e estudar a variação do teor de umidade, especialmente modelos baseados em secagem natural.

A segunda parte da análise é fornecida na Tabela 4. Nesta tabela, apresenta-se o método adotado e a principal contribuição para cada estudo considerado. Pode-se observar que a questão relacionada ao *Machine Learning* e o peso das toras de madeira são os métodos mais utilizados, pois existe uma relação direta entre essas duas variáveis críticas – exatamente a proposição da pesquisa apresentada nesta Dissertação de Mestrado.

É importante destacar que, na literatura científica relevante, não foram encontrados trabalhos abordando métodos de predição da variação de peso – em substituição à variação do teor de umidade – utilizando abordagens de aprendizado de máquina, que é o objetivo deste trabalho. Um resumo do levantamento bibliográfico deste capítulo foi publicado na *International Conference on Enterprise Information Systems – ICEIS*, no ano de 2022. [49].

Tabela 3: Linha do tempo e características dos estudos

ID	Ano	Tipo de madeira	Espécie	País		Objetivo
1	1959	Tora	(não especificado)	Alemanha		Estudar o processo de secagem de madeira
2	1982	Laminada	Carvalho vermelho and Álamo amarelo	Estados Unidos		Estimar o tempo de secagem
3	1985	Tora / Cavaco	(não especificado)	Suécia		Analisar a relação entre a perda de umidade e produção de energia
4	2000	Laminada	Carvalho vermelho do norte, bordo açucarado, American Beech, Yellow Poplar, Pinus Ponderosa e Abeto de Douglas	Estados Unidos		Estimar o tempo de secagem
5	2003	Tora	Pinus Ponderosa e Abeto de Douglas	Estados Unidos		Estimar a perda do teor de umidade por secagem natural em toras com diâmetro pequeno
6	2004	Laminada	Carvalho vermelho do norte, Sugar Maple e Pinus Ponderosa	Estados Unidos		Descrever, em uma equação, o processo de secagem natural
7	2006	Laminada	E. Benthamii, E. Deaneii, E. Dorrigensis, E. Dunni e E. Smithii	Nova Zelândia		O uso potencial de diferentes espécies de eucalipto de acordo com as rachaduras
8	2007	Tora	Giant-Fir e Tsuga	Canadá		Predizer as taxas de secagem em fornos de madeira
9	2009	Tora	(não especificado)	Irlanda		Estimar a taxa de perda de umidade em toras em secagem natural
10	2012	Tora	Pinus	Finlândia		Modelar a mudança do teor de umidade em uma função com variáveis de entrada mensuráveis
11	2012	Tora	Pinus Taeda, Eucalyptus Dunni e Eucalyptus Grandis	Brasil e Estados Unidos		Examinar o impacto do tempo na receita contábil e o custo operacional da perda de peso
12	2012	(não especificado)	Sugi (cedro japonês)	Japão		Predizer o teor de umidade
13	2014	Tora	Pinus Radiata	Nova Zelândia		Observar o efeito causado pelo tempo de armazenamento em toras
14	2015	Lenha	Pinus e Pinus silvestre	Finlândia		Estimar o tempo de armazenamento ideal para pilhas de madeira armazenadas ao ar livre
15	2017	Cavaco	Pinus e Carvalho	França		Predizer o teor de umidade
16	2018	Tora	Pinus Silvestre	Polônia		Avaliar as mudanças em massa e teor de umidade de madeira armazenada em toras
17	2018	Tora	Beech	Polônia		Determinar o período mínimo para que toras de madeira armazenadas em pilhas iniciem o processo de secagem natural
18	2019	Cavaco	Fagus orientalis	Turquia		Predizer o teor de umidade
19	2019	Tora / Laminada	(não especificado)	Canadá		Alocar blocos de corte de madeira para otimizar a produção industrial
20	2019	Tora (árvores)	Eucalyptus Grandis	Brasil		Estimar o volume de madeira em árvores
21	2019	Tora	(Não especificado)	Canadá		Custo econômico da perda de peso da madeira em função do teor de umidade por secagem natural
22	2019	Laminada	Carvalho vermelho, Yellow Poplar e Bordo açucarado	Estados Unidos		Estimar o potencial de uma localização geográfica para secar madeira

Tabela 4: Linha do tempo e principais contribuições das pesquisas

ID	Método	Principal contribuição
1	Estudar a secagem de madeira em fornos de secagem	Estabeleceu temperatura, umidade relativa, velocidade do vento e precipitação como variáveis
2	Deduzir a taxa diária de perda de umidade a partir de dados meteorológicos	Definiu um modelo de regressão para estimar a perda diária do teor de umidade
3	Analisar a umidade e a perda de peso antes e depois do armazenamento	Concluiu que se a madeira armazenada em pilhas é reduzida a cavaco de madeira, a perda de energia pode variar entre 7% e 21% para um período entre 6 e 9 meses de armazenamento
4	Simulação computacional que incluiu o efeito da espessura da madeira no tempo de secagem	Estimou o tempo de secagem natural, até qualquer teor de umidade final, em qualquer dia e em qualquer local onde existam dados de temperatura média e umidade relativa
5	Utilizou dados climáticos, mediu temperatura e umidade relativa a cada 10 minutos	Duas diferentes equações, uma para cada espécie
6	Utilizou registros históricos de temperatura e umidade relativa	Desenvolveu uma equação que permite prever o tempo estimado de secagem natural
7	Mediu as rachaduras e calcula o rendimento em madeira serrada	Expressou a relação entre as rachaduras e o rendimento da madeira e estabeleceu a divisão das toras em classes de diâmetros
8	Aprendizado de máquina com três entradas: teor de umidade inicial, densidade e tempo de secagem	Confirmou a excelente capacidade preditiva deste método de modelagem
9	Pesou continuamente as toras	Estabeleceu o acesso às correntes de vento e a exposição ao sol como determinantes para a estimativa do teor de umidade
10	Mediu o percentual de umidade relativa, temperatura, velocidade do vento e precipitação	Estabeleceu uma equação
11	Pesou constantemente	Demonstrou a diferença de custo industrial entre a madeira verde e toras com 11 semanas de corte
12	Método de aprendizado de máquina utilizando o algoritmo SVM	Proveu a efetividade do método na resolução de problemas deste propósito
13	Pesou as toras a cada 1 e 4 semanas	Demonstrou que toras pequenas, sem qualquer cobertura, secam mais rapidamente, e toras maiores secam mais lentamente
14	Comparou modelos de regressão com um modelo clássico de peso de toras em pilhas	Demonstrou que o peso periódico das toras de madeira é um método efetivo para prever a diferença no teor de umidade
15	Pesou continuamente as toras	Estabeleceu que quanto maiores são as toras armazenadas em pilhas, menor é a perda de umidade
16	Comparou pesos de toras com árvores inteiras	Concluiu que duas semanas é o período mínimo para se ter perda significativa de umidade
17	Relacionou modelos de aprendizado de máquina com um simulador existente	Modelos de aprendizado de máquina atingiram até 94% do ganho máximo teórico
18	Aprendizado de máquina, utilizando o coeficiente de reflexão	Permitiu que esta solução para previsão de teor de umidade seja adequada para implementação direta em processos industriais de madeira para energia
19	Comparou equações de afilamento com modelos baseados em aprendizado de máquina	Métodos de aprendizado de máquina obtiveram resultados iguais ou superiores
20	Pesou continuamente as toras	Uma formulação matemática foi proposta que decide a melhor localização para armazenamento
21	Lógica <i>fuzzy</i> , relacionando o teor de umidade com a temperatura e o tempo de secagem	Os resultados dos modelos <i>fuzzy</i> demonstraram que podem ser uma ferramenta útil para secagem da madeira
22	Simulações	Saídas: a data estimada quando a madeira irá alcançar o teor de umidade-alvo e outros

4 METODOLOGIA

Com a crescente – e contínua – popularização da tecnologia, não tem sido difícil armazenar grandes volumes de dados em computadores e recuperá-los quando necessário. Muito embora os dados armazenados sejam um bem valioso de uma organização, muitas empresas se deparam com o problema de ter “muitos dados, mas pouco conhecimento” (*data rich but knowledge poor*) [37].

Uma das maiores dificuldades que se tem, em função das inerentes limitações humanas, é a capacidade de se analisar e interpretar grandes volumes de dados. Nesse sentido foi proposta, em 1996, a metodologia de *Knowledge Discovery in Databases*, ou *Descoberta de Conhecimento em Bancos de Dados* [25], com o objetivo de justamente preencher esta lacuna. O processo de KDD é composto por um conjunto de etapas contínuas, ditas não-triviais, interativas e iterativas¹, cujo objetivo é identificar padrões que sejam, além de válidos, possíveis de serem compreendidos – além de realmente úteis – com base em uma grande quantidade de dados inicial.

A metodologia original de KDD propõe cinco etapas: seleção dos dados; pré-processamento e limpeza dos dados; transformação dos dados; Mineração de Dados (*Data Mining*); e, por fim, a interpretação e avaliação dos resultados. A Figura 4 apresenta as principais etapas da metodologia original de KDD. Entretanto, ao longo dos anos o processo foi sendo adequado com o surgimento de novos métodos. Entre as adaptações presentes na literatura científica, a mais usual é a combinação do pré-processamento e da transformação como uma única fase [17], conforme mostra a Figura 5.

Nesta proposição, a segunda etapa é o *pré-processamento*, também conhecido por *preparação*. Esta etapa compreende os tratamentos realizados nos dados, independente de eles estarem inconsistentes ou até mesmo ausentes. O pré-processamento é composto de sub-etapas que antecedem a mineração de dados, e tem por objetivo preparar os dados para uma análise eficiente e eficaz, identificando e removendo problemas presentes nos dados antes que os algoritmos de aprendizagem sejam aplicados. De forma geral, o pré-processamento dos dados é um processo semi-automático, pois depende da capacidade

¹ A característica *não-trivial* diz respeito à complexidade existente na execução e manutenção dos processos de KDD; *interativo* porque demanda ação de um cientista ou especialista; e *iterativo* porque permite – e muitas vezes exige – que cada etapa possa ser repetida.

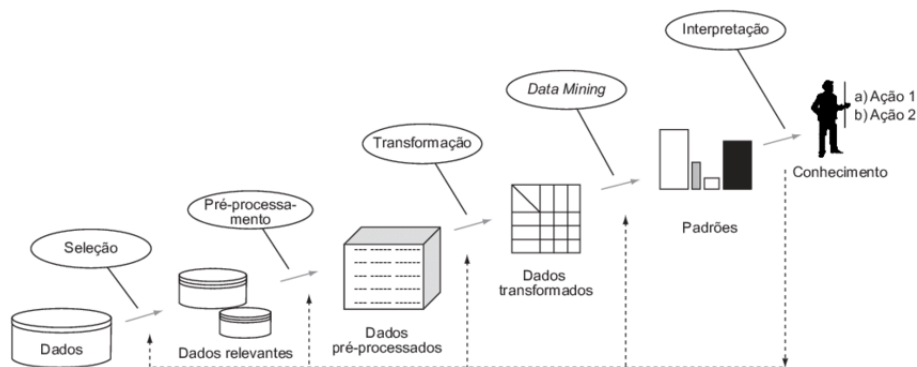


Figura 4: Adaptação do processo de descoberta de conhecimento em bancos de dados. [25]

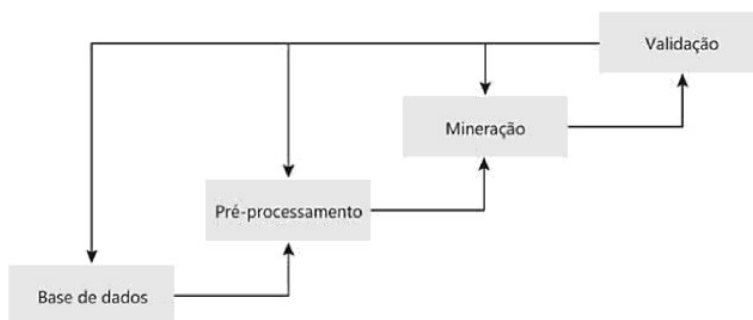


Figura 5: Adaptação da sintetização do processo de descoberta de conhecimento em bancos de dados. [17]

da pessoa que o conduz em identificar os problemas presentes nos dados – bem como a natureza destes problemas – e utilizar métodos mais apropriados para solucionar cada um destes problemas [8]. Por padrão, é a etapa mais extensa e trabalhosa do processo de descoberta de conhecimento. Por outro lado, pode ser uma das mais importantes, pois é nesta etapa que se assegura a qualidade dos fatos que os dados representam. É importante salientar que a qualidade do conhecimento extraído a partir do processo de descoberta de conhecimento é amplamente determinada pela qualidade dos dados de entrada.

Na etapa de pré-processamento estão incluídas a *limpeza* (remoção de possíveis ruídos e inconsistências nos dados), a *integração* (a combinação, quando houver, de dados obtidos a partir de múltiplas fontes), a *seleção* (também conhecida por *redução*, que é a escolha dos dados que serão relevantes à posterior análise) e a *transformação*². Esta pode ser entendida como a consolidação dos dados em formatos apropriados para a mineração, consistindo na aplicação de algumas técnicas – que podem variar de acordo com o problema ou com o próprio formato disponível dos dados, com o objetivo de realizar a procura de atributos dos dados que sejam potencialmente úteis, visando reduzir o número efetivo de atributos a serem considerados pelos algoritmos de mineração da etapa seguinte.

²Na metodologia original de KDD [25], a transformação consistia em uma etapa própria do processo de descoberta de conhecimento, e não estava inserida no contexto do pré-processamento.

A terceira etapa é a *mineração* [83], e consiste na aplicação de algoritmos capazes de extrair conhecimentos a partir dos dados pré-processados. Nesta etapa, é realizada a busca efetiva por conhecimentos úteis que possam emergir dos dados, podendo ser considerada como a etapa central do processo de descoberta de conhecimento. A escolha da técnica – ou *algoritmo* – depende, muitas vezes, do tipo de tarefa de mineração a ser realizada [32].

Esta etapa, na metodologia que será utilizada na presente pesquisa, é composta de duas tarefas distintas: *descritivas*, que ditam as propriedades gerais dos dados (como distribuições de frequência, medidas de centro, média, etc.); e *preditivas*, que fazem inferências a partir dos dados com o objetivo de gerar previsões, verificando determinadas classes de objetos ou estimando determinados valores.

A quarta e última etapa é a *avaliação*, que consiste na validação dos resultados obtidos na etapa de mineração, identificando quais dos conhecimentos são, de fato, verdadeiramente úteis. É nesta etapa que ocorre a interpretação dos padrões que foram minerados, bem como das regras que foram criadas na etapa de mineração. Aqui, é comum existir um ponto de retorno para etapas anteriores do processo de KDD, objetivando a resolução de potenciais conflitos com o conhecimento que foi previamente extraído, ou até mesmo para incorporação deste conhecimento no sistema.

Ainda que o processo de KDD seja subdividido em etapas, é importante ressaltar que elas são, ao mesmo tempo, interdependentes e correlacionadas de forma que suas interações influenciam diretamente o resultado final de descoberta de conhecimento.

Com o objetivo de resolver o problema de pesquisa aplicada na presente dissertação, este trabalho propõe a adoção metodológica do processo de KDD sintetizado acima [17].

4.1 Conjunto de dados

Os dados utilizados para o desenvolvimento dos modelos específicos de predição – objeto de pesquisa do presente trabalho – possuem diferentes origens. Por este motivo, se torna muito importante descrever de forma detalhada o processo de integração de diferentes bases de dados, resultando no conjunto de dados final.

Em se tratando da composição dos dados considerados, pode-se categorizá-los em dois distintos grupos: *dados operacionais* e *dados meteorológicos*.

Os *dados operacionais* são os dados específicos de armazenamento das toras de madeira. Nesta categoria estão presentes informações como a identificação das pilhas nas quais as toras de madeira estão armazenadas, datas iniciais e finais de movimentação de toras em cada pilha, pesos totais das toras movimentadas, tanto de entrada quanto de saída, e dados biológico-florestais das toras, como o tipo da madeira (*pinus* e *eucalipto*), espécie da madeira, comprimento das toras, classe de diâmetro das toras – conforme exigência do mercado internacional, e a presença (ou não) de casca nas toras.

Já os *dados meteorológicos* são os dados atmosféricos que possuem qualquer tipo de

relação – direta ou indireta – com a variação do teor de umidade na madeira. Tais dados são extremamente importantes para a construção dos modelos de predição pois, com base no levantamento bibliográfico realizado na Seção 3, sabe-se que o processo de variação do teor de umidade – e consequente variação do peso das toras – é extensamente baseado, também, em alguns dados meteorológicos específicos.

As duas categorias de dados foram obtidas de origens distintas. Por este motivo, se faz ainda mais importante sua categorização. A seguir, serão detalhadas as formas de obtenção dos dados, bem como sua detalhada descrição e análise descritiva.

4.1.1 Obtenção dos dados

Os *dados operacionais* utilizados neste trabalho de pesquisa foram fornecidos de forma gratuita por uma empresa cuja finalidade é justamente comprar madeira brasileira (seja a madeira já cortada, em toras, ou diretamente ainda nas florestas), na região sul do Brasil, e depois exportar para o mercado europeu e asiático – tendo como principal foco a China. Os pátios de armazenamento de toras da referida empresa, onde ficam as pilhas, estão localizados na cidade do Rio Grande, estado do Rio Grande do Sul, igualmente ao sul do Brasil.

Como a localização exata dos pátios de armazenamento é conhecida, com base nela, foi possível obter todos os dados meteorológicos históricos disponibilizados pela estação de medição específica através do Instituto Nacional de Meteorologia (INMET³) – órgão oficial e diretamente ligado ao Ministério da Agricultura, Pecuária e Abastecimento do Brasil (MAPA)⁴. Buscou-se, então, no sistema Web⁵ do próprio INMET, a estação de medição automática mais próxima dos pátios de armazenamento. Precisamente, os dados foram obtidos da Estação de medição A802⁶, localizada nas coordenadas 32°04'43.7" S 52°10'03.8" W, no período compreendido entre as datas de 11/02/2021 e 21/01/2022.

Para cada pilha de toras de madeira existente em todo o histórico de armazenamento da referida empresa, buscou-se a primeira data em que ocorreu alguma movimentação (o primeiro depósito de toras de madeira na pilha) e também a última data em que ocorreu alguma movimentação (depósito ou retirada de toras da pilha). De posse desse intervalo de datas de movimentação – a primeira data e a última data – buscou-se no banco de dados do INMET os dados meteorológicos históricos contidos no período, para a mesma localização. É importante salientar, também, que os dados meteorológicos são registrados de forma diária e automática pelas estações meteorológicas do INMET.

Entretanto, os equipamentos das estações automáticas de meteorologia, assim como outros equipamentos sujeitos às intempéries, podem apresentar eventuais problemas de funcionamento. Por este motivo, após a obtenção dos dados meteorológicos da referida

³<https://portal.inmet.gov.br/>

⁴<https://www.gov.br/agricultura/pt-br>

⁵<https://bdmep.inmet.gov.br/>

⁶<https://tempo.inmet.gov.br/TabelaEstacoes/A802>

estação automática de medição, observou-se que existiam algumas datas sem qualquer tipo de registro. Como o maior intervalo encontrado sem medições automáticas foi de dezesseis dias (entre 24/02/2021 a 12/03/2021), e levando em conta que tal intervalo não pressupõe uma variação climática e atmosférica relevante, foi definida a média aritmética entre as duas datas anterior e posterior mais próximas para imputar os dados faltantes. Por exemplo, se não houver medição meteorológica realizada no dia 10, mas houver uma medição realizada no dia 9 e também uma medição realizada no dia 11, a medição do dia 10 será preenchida com a média aritmética das medições dos dias 9 e 11. Tal intervenção se mostrou efetiva para mitigar grandes amplitudes nas variações meteorológicas.

Após vencidas estas etapas, os dados *operacionais* e *meteorológicos* foram integrados num mesmo conjunto de dados, final, utilizado na presente pesquisa.

4.1.2 Descrição dos atributos

O conjunto de dados que foi gerado – já integrado – considerado neste trabalho, após realizadas as operações supracitadas, restou em um total de vinte e três atributos diferentes, divididos entre atributos nominais, categóricos e ordinais. Neste conjunto estão contidos todos os dados que foram possíveis de extrair das diferentes origens, categorizados em *dados operacionais* e *dados meteorológicos*, conforme explicados anteriormente.

A seguir, na Tabela 5, é mostrada a relação de todos os atributos do conjunto de dados final e, para cada um deles, são apresentadas informações úteis e complementares, com o objetivo de propiciar um melhor entendimento, como o tipo de dado, valor mínimo, máximo e médio (moda para categóricos e mediana para ordinais), além da descrição.

É possível observar que a maioria dos atributos presentes no conjunto final de dados são do tipo numérico contínuo (60%). Oito atributos são do tipo categórico (35%) e, ainda, um atributo é numérico discreto. Neste ponto, cabe destacar que o atributo presente no conjunto de dados e que irá originar a criação do atributo-alvo deste trabalho, ou seja, do atributo a ser predito, é o atributo *estorno*. Este atributo representa a relação direta da variação do teor de umidade através do peso da madeira⁷.

Por fim, se faz importante destacar que, em relação ao número de instâncias, o conjunto final de dados está composto por 759 (setecentos e cinquenta e nove) casos.

4.1.3 Estatística descritiva dos dados

De forma a aprofundar o conhecimento sobre os dados, é importante realizar uma análise descritiva sobre os atributos que compõem o conjunto de dados. Por isso, foram plotados gráficos com as frequências dos atributos, buscando uma melhor compreensão

⁷O atributo *estorno* é o resultado de uma opção, presente no *software* da empresa que cedeu os dados, que foi desenvolvida para que o setor de controle operacional da empresa conseguisse deduzir a perda de água (resultante da perda de umidade) do estoque de madeira presente na pilha. Quando existe saldo positivo de estoque teórico, mas este saldo não existe fisicamente no pátio, essa diferença é entendida pela empresa como sendo *perda de umidade*, e é cadastrado um estorno nessa mesma quantidade.

Tabela 5: Detalhamento dos atributos presentes no conjunto final de dados.

Atributo	Tipo	Min	Max	Med/Moda	Descrição
PILHA	Categórico	NA	NA	NA	Identificador da pilha de armazenamento de toras de madeira
DATA_INICIO	Data	11/02/2021	17/10/2022	12/11/2021	Data de início dos depósitos de toras na pilha
DATA_FIM	Data	15/03/2021	26/10/2022	21/01/2022	Data de término dos depósitos e/ou retiradas de toras da pilha
PRODUTO	Categórico	NA	NA	EUCALYPTUS	Tipo de madeira presente na pilha (<i>pinus</i> ou <i>eucalipto</i>)
ESPECIE	Categórico	NA	NA	GRANDIS	Espécie da madeira presente na pilha
DIAMETRO	Ordinal	8 a 12	40 a 60	20 a 30	Classe de diâmetro da seção transversal, medida pela extremidade mais fina das toras de madeira na pilha
COMPRIMENTO	Ordinal	3,1	6	5,2	Classe de comprimento (tamanho, em metros) das toras de madeira
CASCA	Categórico	NA	NA	YES	Se as toras de madeira da pilha possuem ou não possuem casca
TEMP_MIN	Contínuo	1,9	24	5,4	Temperatura mínima (em °C) do período
TEMP_MAX	Contínuo	12,7	39,4	28,5	Temperatura máxima (em °C) do período
TEMP_MED	Contínuo	7,73	37,7	17,2	Temperatura média (em °C) do período
TEMP_PO	Contínuo	4,2	29,4	12,6	Temperatura média (em °C) do período na qual a água presente no ar ambiente passou ao estado líquido
PRECIPITACAO	Contínuo	0	1.156	95,4	Precipitação total (em milímetros) do período
PATM_MED	Contínuo	1.005	1.535,4	1.016,4	Pressão atmosférica média (em hPA) do período
UR_MIN	Contínuo	21	91	32	Umidade relativa mínima do ar (em %) do período
UR_MED	Contínuo	59,5	117,5	76,4	Umidade relativa média do ar (em %) do período
VENTO_MED	Contínuo	1,1	5,4	3	Velocidade média do vento (em m/s) no período
RAJ_MAX_MED	Contínuo	3,7	19,3	10,3	Velocidade de rajada de vento máxima média (em m/s) do período
QTD_DIAS	Discreto	0	425	47	Número de dias do período
TOT_ENTR	Contínuo	940	8.128.820	216.740	Total de entradas (em Kg) de toras na pilha
TOT_SAID	Contínuo	0	6.628.260	168.480	Total de saídas (em Kg) de toras da pilha
DIFERENCA	Contínuo	-169.280	1.500.560	18.990	Diferença (em Kg) entre o total de entradas e o total de saídas
ESTORNO	Contínuo	0	1.500.560	9.220	Quantidade estornada (em Kg) da pilha – em teoria, este atributo deveria estar diretamente relacionado à perda de água

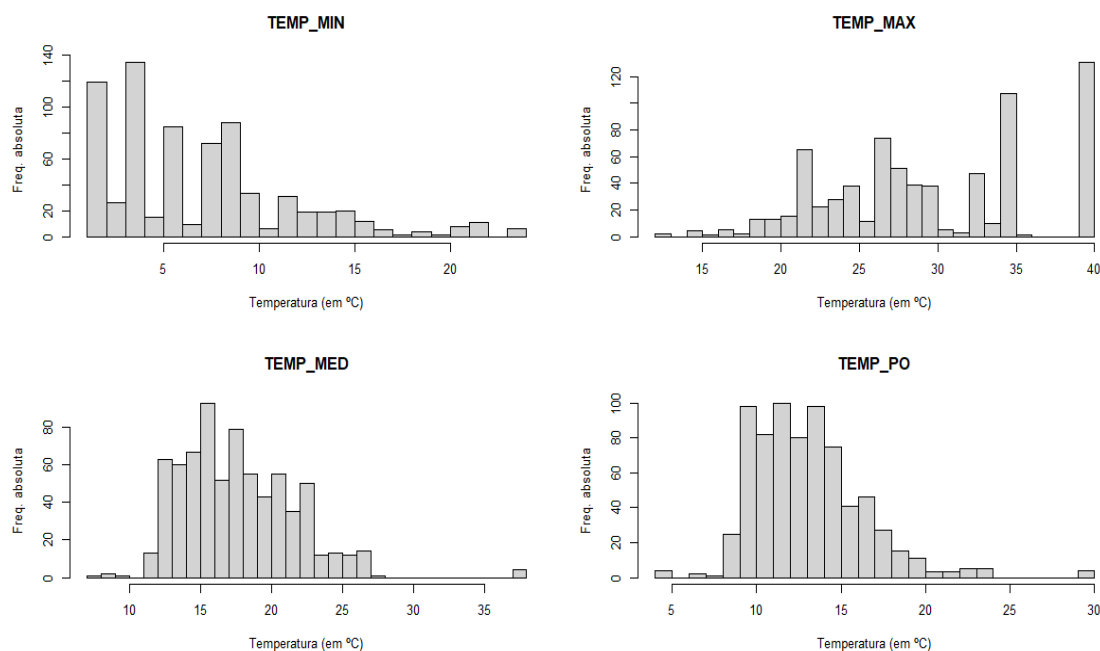


Figura 6: Distribuições das temperaturas durante o período observado de armazenamento das toras de madeira.

visual das distribuições de valores de cada atributo a ser analisado.

Inicia-se com as distribuições das temperaturas (mínima, máxima, média e em ponto de orvalho⁸), obtidas dos *dados meteorológicos* e presentes no conjunto de dados, observadas na Figura 6. Nestes gráficos, pode-se observar que, durante o período de armazenamento das toras de madeira analisado na presente pesquisa, havia predominância de dias com temperaturas mais amenas (média de 17,2 °C). Ao mesmo tempo, observa-se uma grande amplitude de temperaturas durante o mesmo período, com variações entre os dois extremos de temperatura (mínima e máxima).

Com relação às distribuições da precipitação total, da pressão atmosférica média e das umidades relativas – mínima e média – do período de armazenamento das toras na pilha (Figura 7), depreende-se que há predominância de períodos mais secos, sem precipitação capturada pela estação meteorológica (0 mm); a pressão atmosférica média do período é praticamente constante (1000 hPA), devido à localização geográfica dos pátios de toras e da estação meteorológica ser ao nível do mar; já nas umidades relativas, há predominância de reduzidas umidades nas medições mínimas (20%), tendo nas medições médias a umidade predominante entre 70% e 80%.

Continuando a análise descritiva, são apresentadas as distribuições da velocidade média do vento e das médias de rajada máxima de vento (Figura 8). Observa-se que, durante o período de armazenamento observado, a velocidade média do vento predomi-

⁸Temperatura do ponto de orvalho é a temperatura até a qual o ar deve ser resfriado para que a condensação de água se inicie, ou seja, para que o ar fique saturado de vapor de água. Na temperatura do ponto de orvalho, a quantidade de vapor de água presente no ar é máxima.

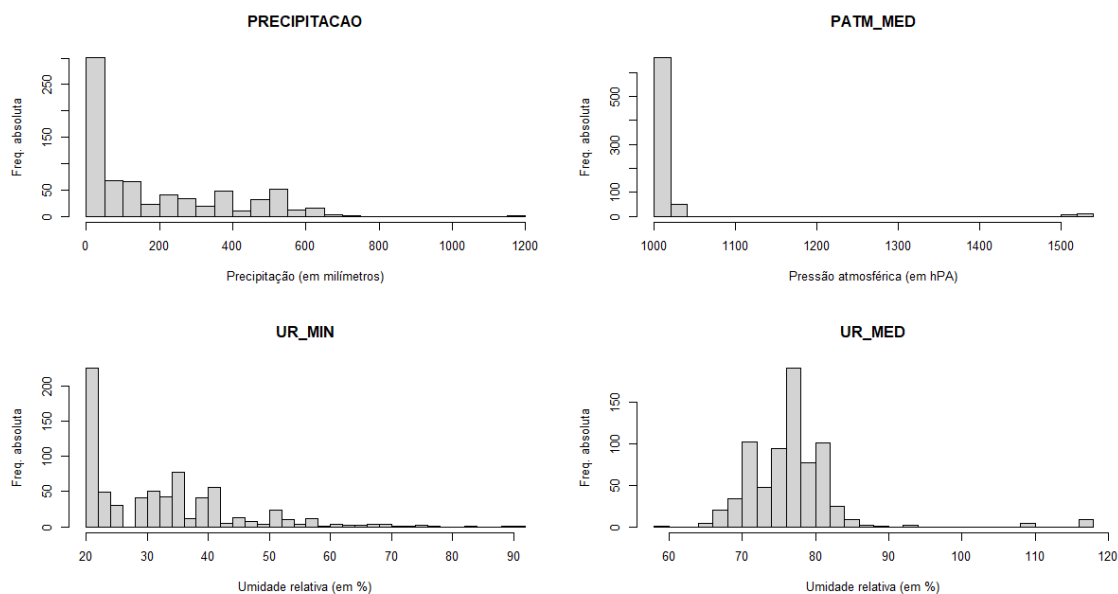


Figura 7: Distribuições de precipitação total, pressão atmosférica média, umidade relativa mínima e umidade relativa média durante o período observado de armazenamento das toras de madeira.

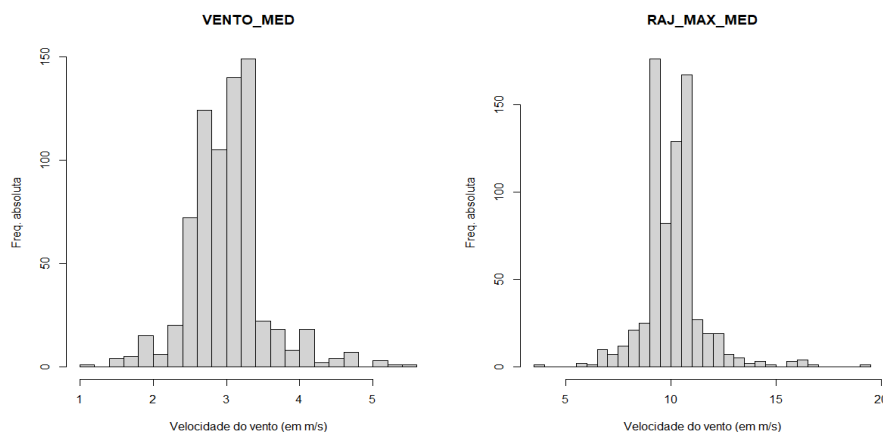


Figura 8: Distribuições de velocidade média do vento, e das rajadas máximas médias do vento durante o período observado de armazenamento das toras de madeira.

nou entre 2,5 m/s (velocidade equivalente a 9 Km/h) e 3,5 m/s (equivalente a 12,6 Km/h), podendo ser observadas as máximas nas rajadas médias entre 7,5 m/s (equivalente a 27 Km/h) e 12,5 m/s (equivalente a 45 Km/h).

Por último, é apresentada a distribuição da quantidade de dias que as toras permaneceram armazenadas nas pilhas (Figura 9). Neste ponto, pode-se verificar facilmente que há um problema na origem dos dados, pois a grande maioria das instâncias não chega a possuir um dia de armazenamento – situação que não ocorre na prática operacional e que é um problema a ser tratado antes da etapa de mineração dos dados, ou seja, na etapa de pré-processamento, seguindo a metodologia adotada neste trabalho de pesquisa aplicada.

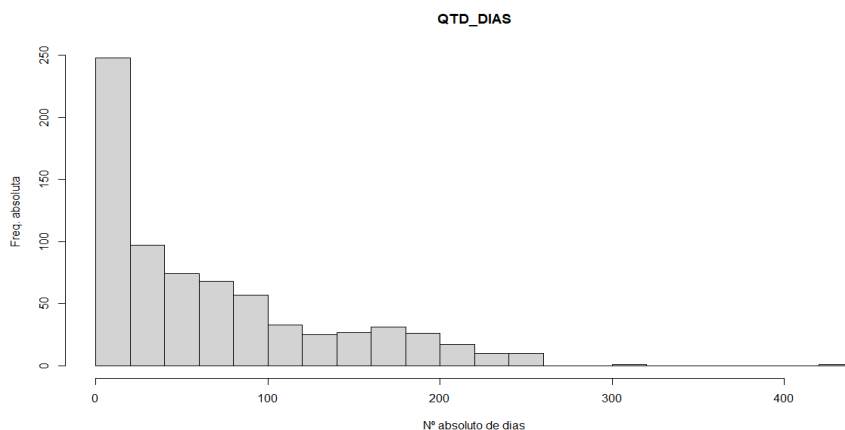


Figura 9: Distribuições da quantidade de dias em que as toras estiveram armazenadas nas pilhas durante o período observado.

Ao verificar com o especialista, a informação obtida foi de que alguns lançamentos foram efetuados no *software* somente para fins de registro. Muitas cargas de toras de madeira foram enviadas diretamente das florestas (origem) para os navios (destino), em operações portuárias de exportação de madeira. Por esse motivo, tais toras de madeira realmente não chegaram a ficar armazenadas nas pilhas. Nestes casos, obviamente, não houve verificação de diferença de peso e, por isso, estas instâncias precisam ser removidas para que os eventuais modelos gerados não se tornem enviesados através destes ruídos.

4.1.4 Pré-processamento dos dados

Dando início formal à etapa de pré-processamento dos dados, foi criado um novo atributo no conjunto de dados, denominado *PERCENTUAL*. Ele representa o percentual de perda de peso de madeira (quantidade estornada) em relação ao somatório de entradas de madeira em cada pilha, podendo ser calculado conforme a seguinte equação:

$$P = (E \times 100) \times T^{-1},$$

onde P representa o percentual de perda de peso, E representa o peso estornado de cada pilha e T representa o somatório dos pesos de entrada de madeira de cada pilha.

Após a criação do atributo supracitado, foram removidas as instâncias que representam pilhas de armazenamento de toras das quais não havia registro de saída de madeira, ou seja, somente entrada. Em teoria, tais pilhas continuavam a receber cargas de madeira ou ainda não estavam finalizadas no momento da obtenção dos *dados operacionais*. Podem representar, ainda, erros de lançamento de dados no *software* pelo setor de controle operacional da empresa responsável pelo fornecimento dos dados.

Também foram removidas as instâncias que não tiveram registro de cargas estornadas, ou seja, instâncias que representam pilhas de armazenamento de toras que, em teoria,

não tiveram perda de peso por umidade registrada. Tais instâncias necessitam ser removidas porque, tanto com base no conhecimento científico adquirido no levantamento bibliográfico realizado na Seção 3, quanto no conhecimento do especialista, em madeira armazenada ao ar livre (tecnicamente, chama-se este processo de *secagem natural*), nas condições dos dados capturados, e sendo as toras de madeira destinadas à exportação, sempre haverá variação no peso das mesmas causado, diretamente, pela diferença de umidade.

Após a remoção destas instâncias, foi iniciado o processo de seleção de características, de forma a remover vieses dos dados para ficarem de acordo com o tipo de aprendizado supervisionado a que se propõe este trabalho – a classificação.

Em situações do mundo real, em que os dados normalmente não estão disponíveis no formato ideal para se iniciar o processo de descoberta de conhecimento – e muitas vezes são obtidos de origens distintas, como é o caso deste trabalho, faz-se necessária a utilização de ferramentas que tornem a mineração de dados mais eficaz.

Nesse sentido, a seleção de características, ou *feature selection*, se torna fundamental, pois pode acelerar os próprios algoritmos que se utilizam de métodos de *Machine Learning*, melhorando a acurácia preditiva e aumentando sua compreensão [45]. Durante o desenvolvimento de um modelo preditivo, reduzir o número de atributos de entrada pode tanto reduzir o custo computacional quanto melhorar a performance preditiva do modelo.

Alguns algoritmos de aprendizado de máquina realizam, de forma intrínseca, algum tipo de seleção de características – tal como o *Random Forest – RF* [13]⁹.

Na revisão bibliográfica apresentada, referente aos trabalhos relacionados (Seção 3), é demonstrado que a literatura científica já realizou, de certa forma, um processo manual de seleção de características – ao determinar que somente alguns dados meteorológicos possuem relação significativa com a variação do teor de umidade. Por esta razão, mais uma vez a revisão bibliográfica se torna importante, pois a produção de conhecimento gerado e acumulado durante a revisão, ainda que de forma não-sistemática, torna o desenvolvimento dos modelos mais otimizado.

Então, para o desenvolvimento dos modelos de predição, dando continuidade à seleção de características, foram removidos os atributos que representam a identificação de cada pilha de toras de madeira (*ID_PILHA*); a quantidade total de madeira retirada de cada pilha (*TOT_SAID*); a diferença entre o peso total de madeira depositado e o peso total de madeira retirado de cada pilha (*DIFERENCA*); as datas que representam o início e o fim da movimentação de madeira em cada pilha (*DATA_INICIO* e *DATA_FIM*); e, por fim, a quantidade de madeira que foi estornada pela empresa (*ESTORNO*).

Nesse momento, se faz necessária a criação do atributo-alvo deste trabalho de pesquisa, que será chamado de *INTERVALO_PERDA*. Este atributo irá representar os intervalos de percentuais de perda de peso de cada pilha de madeira, sempre em relação ao

⁹Este algoritmos está no rol dos que foram utilizados para a resolução do presente problema.

total de entradas.

Para a criação deste atributo no conjunto de dados, se faz necessária uma análise descritiva do atributo *PERCENTUAL* – que também foi criado durante a presente etapa. Para isso, foram realizadas análises distintas, iniciando por uma visual: através de um *boxplot* e de um histograma com diferentes percentuais de perda de peso (Figura 10).

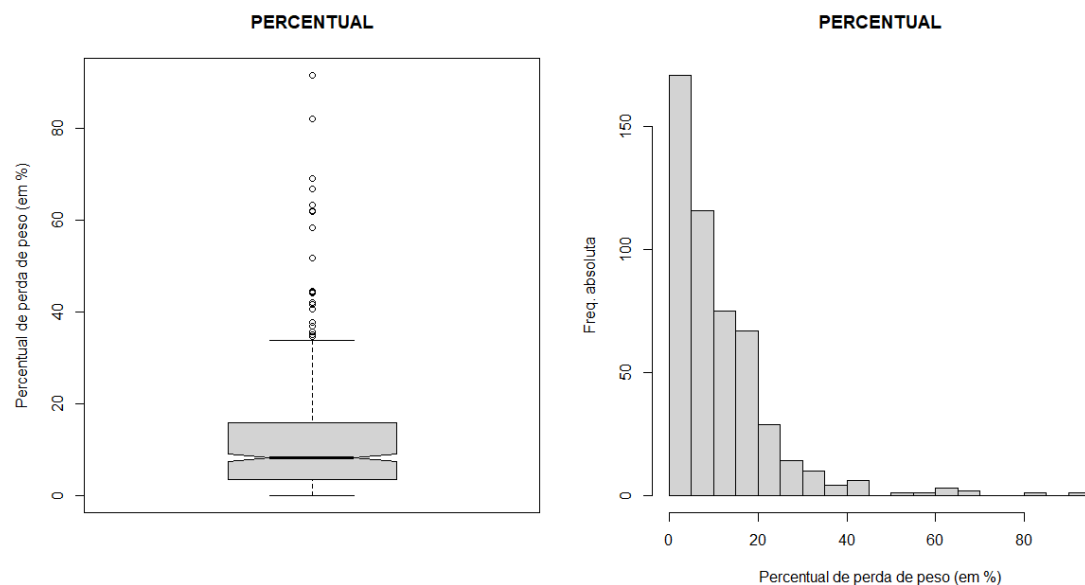


Figura 10: À esquerda, o *boxplot* com os percentuais de perda de peso. À direita, o histograma de frequência das instâncias com diferentes percentuais de perda de peso.

Em seguida são apresentadas as estatísticas básicas do atributo *PERCENTUAL*:

- Valor mínimo: 0,006
- Primeiro quartil: 3,457
- Mediana: 8,186
- Média: 11,422
- Terceiro quartil: 15,803
- Valor máximo: 91,583

Através desta análise descritiva, percebe-se que existem possíveis erros de lançamento no *software* da empresa que originou os dados, pois há registro de percentuais de perda de peso próximos de 100% – o que é fisicamente impossível. Também existem valores de 0%, ou seja, mesma situação. A análise através dos dados que compõem o *boxplot* permite que seja identificado que valores de percentual abaixo do limite inferior 0,005% e acima do limite superior 33,901% podem ser considerados *outliers*, ou seja, instâncias que

se forem mantidas podem gerar ruído no modelo de classificação que será desenvolvido nesta pesquisa.

Um *boxplot* é uma representação visual contendo informações sobre a distribuição dos dados, e é construído sobre cinco pontos-chave: o limite inferior, o primeiro quartil, o segundo quartil (ou mediana), o terceiro quartil e o limite superior. Em termos estatísticos, são considerados *outliers* todas as instâncias que ficam localizadas além dos limites inferior e superior, ou seja, instâncias diferentes da maioria dos dados.

Os limites, tanto superior quanto inferior, podem ser expressos matematicamente pelas equações:

$$LimSup = Q3 + (1,5 \times (Q3 - Q1))$$

$$LimInf = Q1 - (1,5 \times (Q3 - Q1))$$

onde *LimSup* representa o limite superior e *LimInf* o limite inferior do *boxplot*, *Q3* representa o terceiro quartil, e *Q1* o primeiro quartil.

Os limites de um *boxplot* possibilitam uma análise visual e representativa da dispersão dos dados e dos possíveis valores-limite, mas a interpretação e a definição dos limites pode variar de acordo com o contexto. Existem diferentes abordagens para calcular e definir os limites e, dependendo da abordagem utilizada, os resultados podem vir a ser diferentes.

Considerando, neste trabalho, que o conjunto de dados que se têm à disposição não é volumoso; e que instâncias com variação acima de 33,9% são habituais em se tratando de armazenamento de madeira ao ar livre, serão consideradas somente as instâncias com perdas de peso abaixo de 40%, removendo todas as demais. Após a remoção, tem-se as estatísticas abaixo:

- Valor mínimo: 0,006
- Primeiro quartil: 3,246
- Mediana: 7,670
- Média: 9,981
- Terceiro quartil: 15,032
- Valor máximo: 37,756

Após este tratamento, pode-se retomar a criação do atributo *INTERVALO_PERDA*, estabelecendo e preenchendo as instâncias com diferentes intervalos de percentuais de perda, conforme discretização apresentada na Tabela 6.

Tabela 6: Divisão de intervalos de perda de peso.

Intervalo	Percentual de perda de peso
[0-10]	$0 \leq x \leq 10$
(10-20]	$10 < x \leq 20$
(20-30]	$20 < x \leq 30$
(30-inf]	$x \geq 30$

Enfim, depois de estabelecidos os intervalos de perda de peso (em percentual), para fins de desenvolvimento dos modelos de classificação, é preciso remover o atributo *PERCENTUAL* do conjunto de dados – um atributo está fortemente correlacionado com o outro, afinal, um foi estabelecido e preenchido com base nos valores do outro.

Removido o atributo e definida a classe (o atributo-alvo, o que se deseja prever – o atributo *INTERVALO_PERDA*), é momento de verificar o balanceamento das classes no conjunto de dados. A Tabela 7 demonstra a situação presente nos dados, evidenciando o número de instâncias em cada classe a ser predita.

Tabela 7: Balanceamento entre as classes do conjunto de dados.

Intervalo	Nº de instâncias	Proporção
[0-10]	288	59,1%
(10-20]	142	29,2%
(20-30]	43	8,8%
(30-inf]	14	2,9%
Total	487	100,0%

Após análise dos dados acima, pode-se observar com clareza que o maior número de instâncias do conjunto de dados está representado nas classes [0-10] e (10-20], ou seja, a grande maioria das pilhas presentes no conjunto de dados fornecido pela empresa possui uma perda de peso (em razão da perda do teor de umidade), em percentual, menor do que 20%. Para verificar a integridade desta informação, pode-se comparar a perda de peso com a quantidade de dias das toras armazenadas nas pilhas. Afinal, quanto maior o tempo armazenado, em teoria, maior será a perda de peso – até a variação ficar próxima a zero, com o teor de umidade presente nas toras ser o mínimo possível.

Ao verificar as estatísticas descritivas de todas as instâncias no atributo *QTD_DIAS*, tem-se abaixo:

- Valor mínimo: 0,00
- Primeiro quartil: 22,00
- Mediana: 68,00
- Média: 81,26
- Terceiro quartil: 127,00

- Valor máximo: 425,00

Ao analisar o atributo *QTD_DIAS*, pode-se perceber que há uma observação com 425 dias de armazenamento, podendo ser considerada um *outlier*, pois ultrapassa o limite superior de 284,5 dias. Após a remoção da referida instância, as novas estatísticas descritivas são:

- Valor mínimo: 0,00
- Primeiro quartil: 22,00
- Mediana: 68,00
- Média: 80,56
- Terceiro quartil: 126,00
- Valor máximo: 255,00

Considerando os dados acima, então, que em se tratando de perda do teor de umidade em madeira armazenada ao ar livre (*natural drying*) 20% é um percentual pequeno, pode-se depreender que as toras de madeira observadas neste conjunto de dados, em razão da grande quantidade média de dias que permaneceram armazenadas nas pilhas (praticamente 81 dias) já haviam sido cortadas há mais tempo – informação descoberta de análise descritiva dos dados e que é importante para a posterior análise dos modelos preditivos.

O passo seguinte é o particionamento do conjunto de dados. O particionamento aleatório em conjuntos de treinamento e teste foi um padrão estabelecido pelo aprendizado de máquina por décadas e baseia-se na suposição que, ao fixar o conjunto de dados de forma estática, se torna possível avaliar diferentes algoritmos ou modelos de aprendizado de máquina [82].

Num particionamento de dados, o comportamento esperado é de que as futuras partições reproduzam a distribuição geral do conjunto de dados. Para esta pesquisa, 80% das instâncias foram selecionadas para o treinamento e validação dos modelos, e os 20% restantes foram utilizados para teste dos modelos.

O pré-processamento relacionado ao balanceamento das classes será considerado na próxima seção, ao apresentar as múltiplas configurações de modelos a serem ajustados.

4.2 Mineração dos dados

Neste trabalho foram treinados diferentes modelos de classificação de dados, utilizando diferentes algoritmos, com configurações distintas, os quais foram aplicados sobre

subconjuntos distintos dos mesmos dados. Esta seção tem por objetivo apresentar a metodologia de sintonia dos hiperparâmetros destes algoritmos, bem como especificar as estratégias de amostragem aplicadas.

Os algoritmos considerados para o desenvolvimento do modelo que se propõe a resolver o presente problema de pesquisa foram: *Random Forest* (RF), *Logistic Regression* (LR), *Support Vector Machine* (SVM), *Gaussian Naive Bayes* (NB), *Adaboost* (AB), *Gradient Boosting* (GB) e *Multilayer Perceptron* (MLP). Detalhes dos algoritmos podem ser consultados na Seção 2.4. Para implementação dos modelos, foi utilizada a biblioteca de aprendizado de máquina *scikit-learn*¹⁰ [60].

Em relação à configuração dos experimentos, estes algoritmos foram aplicados sobre os dados de treinamento de duas formas distintas: com os valores *default* de seus hiperparâmetros na biblioteca de aprendizado de máquina utilizada, e com seus hiperparâmetros ajustados ou sintonizados (*tuned*). Em relação aos dados de treinamento, essas duas configurações foram aplicadas sobre dados diferentes: num primeiro momento, sobre os dados desbalanceados, ou seja, da forma com que foram obtidos e pré-processados; após, sobre os dados balanceados (com instâncias reais e sintéticas, criadas artificialmente, igualando as classes minoritárias à classe majoritária, garantindo o mesmo número de amostras em todas as classes).

Em seguida, foi escolhida arbitrariamente uma lista de 10 (dez) sementes (*seeds*), usadas para inicializar um gerador de números pseudoaleatórios¹¹, para emular uma amostra de tamanho 10 – na qual, cada *seed* funciona como um representante da amostra. Dessa forma, foi gerado um modelo diferente para cada combinação de *seed*, algoritmo de classificação, configuração de hiperparâmetros (*default* e *tuned*), e conjunto de dados de treinamento (desbalanceado e balanceado). Os modelos foram avaliados sobre os dados de teste e comparados com base na métrica μF_1 , ou seja, na média ponderada das medidas F, considerando a quantidade de instâncias por classe. Conforme apresentado na fundamentação teórica (Seção 2.3), a F1 é a média harmônica entre a *precisão* e a *revocação* – com peso igual para ambas as medidas. Dessa forma, os modelos devem ser tão bons para identificar a classe correta quanto para não considerar falsos positivos nos resultados das predições.

Para encontrar a melhor configuração possível dos hiperparâmetros (*tuned*) em cada experimento, foi utilizado o algoritmo *Grid Search Cross-Validation* – *GridSearchCV* [47]. Este algoritmo recebe como entrada um conjunto de valores para cada hiperparâmetro, definidos pelo especialista. A este conjunto, chama-se *espaço de busca*. Além do espaço de busca, é preciso definir a métrica de avaliação a ser considerada e o número de *folds* a serem avaliados por validação cruzada. Após estes ajustes, é realizada a busca

¹⁰Disponível em <https://scikit-learn.org>

¹¹Ao idealizar que os resultados possuam um comportamento aleatório, mas igual em todas as vezes em que serão testados/executados durante o trabalho, é preciso arbitrar uma *seed*

pelo *best tuning*, combinando todos os possíveis valores de cada hiperparâmetro presente no espaço de busca. Ao final da execução, é possível encontrar a melhor configuração dos hiperparâmetros, ou seja, o melhor modelo possível – dada a métrica escolhida – ajustado aos dados de treinamento. Também é possível analisar o *ranking* com todos os modelos ajustados, permitindo comparações entre custo computacional de treinamento e eficácia de cada modelo.

Na Tabela 8, todas as possibilidades de hiperparâmetros testados para cada algoritmo são evidenciados no espaço de busca arbitrado para este trabalho, onde:

n_estimators (RF) Número de árvores de decisão que serão construídas no modelo e que irão compor a floresta. Quanto maiores os valores deste hiperparâmetro, potencialmente menos vieses terá o modelo de classificação a ser treinado – e mais robusto tende a ser o modelo. Em contrapartida, quanto mais altos os valores, maior será o custo computacional para o treinamento do modelo;

max_features (RF) Número máximo de características que serão levadas em consideração na divisão de cada árvore em cada nó. Valores menores deste hiperparâmetro podem melhorar a capacidade de generalização do modelo, reduzindo a correlação entre as diferentes árvores de decisão na floresta. Já os valores maiores permitirão que as árvores tenham mais características à disposição, podendo considerar mais informações presentes nos dados. O valor *sqrt* significa o valor inteiro mais próximo da raiz quadrada do número total de características, enquanto o valor *log2* significa o valor inteiro mais próximo do logaritmo base 2 do número total de características.

max_depth (RF) Profundidade máxima que cada árvore da floresta pode ter, ou seja, este hiperparâmetro regula o número máximo de divisões que podem ser feitas ao longo de um caminho desde a raiz até as folhas da árvore. É importante que a profundidade de cada árvore seja regulada para evitar que a árvore se ajuste muito bem aos dados específicos de treinamento e, neste caso, não sendo capaz de generalizar o modelo para novos dados desconhecidos. A limitação da profundidade das árvores também influencia o custo computacional no tempo de treinamento do modelo.

max_leaf_nodes (RF) As árvores de decisão crescem dividindo os nós em subconjuntos, com base nas características presentes nos dados de treinamento. A divisão continua iterativamente até que o número máximo de nós que representam pontos de parada seja encontrado, ou seja, até que o número de nós em que não ocorrem mais divisões ou ramificações da árvore seja atingido. Estes nós representam os pontos nos quais os rótulos são atribuídos. Igualmente ao *max_depth*, a limitação deste hiperparâmetro ajuda a evitar *overfitting* do modelo sobre os dados de treinamento.

C (LR) Valor que controla a intensidade da força de regularização. Quanto menor o valor de C, maior a penalização dos coeficientes – e mais próximos de zero eles estarão. Isso significa que o modelo poderá ser menos preciso, mas com pequenas chances de *overfitting*. Por outro lado, valores maiores de C permitem que os coeficientes sejam mais flexíveis, adaptando-se aos dados de treinamento e aumentando o custo computacional, mas tornando o modelo mais suscetível ao *overfitting*.

max_iter (LR) Número máximo de iterações que serão permitidas durante o treinamento do modelo. No processo de treinamento, o algoritmo busca encontrar os coeficientes que melhor se ajustam aos dados de treinamento para minimizar a função de perda. Cada iteração significa uma passagem completa pelos dados de treinamento durante a otimização dos coeficientes.

multi_class (LR) Hiperparâmetro que especifica qual será a estratégia escolhida pelo algoritmo para lidar com problemas multiclasse. A opção *multinomial* estende o modelo para prever diretamente a probabilidade de a instância pertencer a cada classe, considerando todas as classes simultaneamente. A opção *ovr* (*one-versus-rest*) considera uma classificação binária para cada classe individual, considerando cada classe como positiva e todas as demais como negativas. Por último, a opção *auto* seleciona automaticamente a estratégia, dependendo da disposição dos dados de treinamento.

kernel (SVM) Tipo de função-núcleo a ser utilizada pelo algoritmo na transformação do espaço a ser utilizado. O *kernel* permite a busca do melhor hiperplano de separação entre as classes presentes nos dados. A opção *linear* utiliza uma transformação linear nos dados de entrada, tendo maior eficácia quando os dados são linearmente separáveis.

C (SVM) Número que representa o parâmetro de penalidade que será aplicado ao erro no modelo, como uma espécie de suavização da margem de separação das classes. Valores mais altos de C indicam maior penalidade, significando que o modelo irá tentar classificar corretamente o maior número possível de instâncias – diminuindo a margem de separação entre as classes, podendo ocasionar *overfitting*. Valores menores indicam menos penalidades, permitindo algumas classificações incorretas.

max_iter (SVM) Número máximo de iterações a ser realizado pelo algoritmo no ajuste dos pesos – tarefa realizada durante a busca dos melhores pesos dos vetores de suporte – independente se o algoritmo atingiu, ou não, a convergência.

var_smoothing (NB) Número adicionado à variância de cada variável de entrada, para evitar problemas numéricos, quando a variância for igual a zero (ou muito próxima de zero). Este número é responsável por uma espécie de suavização da variância,

evitando divisões por zero durante o cálculo das probabilidades condicionais, já que este algoritmo assume que as variáveis de entrada estão dispostas em uma distribuição normal e utiliza, para realizar as predições, as médias e as variâncias das variáveis.

n_estimators (AB) Número máximo de *weak learners* que serão treinados e combinados para formar o modelo final. Números grandes definidos neste hiperparâmetro podem levar a modelos mais robustos, mas com maior possibilidade de *overfitting*.

algorithm (AB) Hiperparâmetro que especifica o algoritmo-base a ser utilizado como *weak learner* durante o treinamento do modelo. A opção *SAMME.R* (*Stagewise Additive Modeling using a Multiclass Exponential loss function, Real version*) é mais utilizada em problemas de classificação multiclasse, ajustando os pesos dos *weak learners* de acordo com a probabilidade de classificação correta, minimizando a função de perda. Já a opção *SAMME* (*Stagewise Additive Modelling using a Multiclass Exponential loss function*) ajusta os pesos com base na classificação de cada instância, sendo ela correta ou incorreta.

learning_rate (GB) Hiperparâmetro que controla a contribuição de cada *weak learner* individual para o modelo, considerando que a cada iteração do algoritmo, um novo *weak learner* é treinado para corrigir os erros residuais cometidos pelos *weak learners* anteriores. É um hiperparâmetro importante a ser considerado, pois ele pode ajustar tanto a taxa de aprendizado de cada *weak learner* como o equilíbrio entre o ajuste aos dados de treinamento e a sua capacidade de generalização. Usualmente, os valores comuns para este hiperparâmetro são menores que 1, como 0,1 ou 0,01.

max_depth (GB) Igualmente ao *Random Forest*, este hiperparâmetro controla a profundidade máxima de cada árvore de decisão individual a ser combinada na construção do modelo final, ou seja, o número máximo de divisões que podem ser feitas ao longo do caminho entre a raiz e as folhas das árvores.

n_estimators (GB) Igualmente ao *Adaboost*, este hiperparâmetro define o número de *weak learners* que serão treinados e combinados para o desenvolvimento do modelo final.

max_iter (MLP) Número máximo de vezes que o algoritmo irá percorrer todo o conjunto de treinamento, buscando otimizar o ajuste dos pesos. Cada iteração corresponde a um passo no processo de ajuste, levando em consideração o algoritmo de otimização escolhido.

hidden_layer_sizes (MLP) Hiperparâmetro que permite especificar o número de *perceptrons* presentes em cada camada oculta da rede neural. Este hiperparâmetro aceita

conjuntos de valores inteiros, também, como entrada, correspondendo ao número de *perceptrons* em cada camada oculta específica. Como exemplo, a tupla "(150, 100, 50)" indica três camadas ocultas, uma com 150, outra com 100, e outra com 50 *perceptrons*.

activation (MLP) Hiperparâmetro que permite escolher entre diferentes funções de ativação a ser considerada nos *perceptrons* das camadas ocultas ou na camada de saída da rede neural. A opção *tanh* aplica uma função tangente hiperbólica aos valores de entrada, mapeando os valores para o intervalo entre -1 e 1. A opção *relu* aplica uma função que retorna o valor de entrada se for positivo, e 0 se não for positivo. Por último, a opção *logistic* aplica a função logística aos valores de entrada, mapeando os valores para o intervalo entre 0 e 1, sendo mais aplicada a problemas de classificação binária.

solver (MLP) Hiperparâmetro que define qual será o algoritmo de otimização, a ser utilizado pelo modelo, para ajustar os pesos e minimizar a função de perda da rede neural durante o treinamento. A opção *sgd* (*Stochastic Gradient Descent*) atualiza os pesos em pequenos lotes de dados de treinamento de forma iterativa, sendo adequado para grandes volumes de dados. A opção *adam* (*Adaptive Moment Estimation*) adapta a taxa de aprendizado de acordo com as características dos gradientes anteriores. Já a opção *lbfgs* (*Limited-memory Broyden-Fletcher-Goldfarb-Shanno*) utiliza o método Quasi-Newton, com aproximação da matriz Hessiana utilizando memória limitada, sendo mais apropriado para pequenos volumes de dados.

learning_rate (MLP) Hiperparâmetro que define o tamanho do ajuste que é realizado para atualizar os pesos da rede neural durante o processo de otimização. Valores maiores de *learning_rate* atualizarão os pesos em incrementos maiores, ocasionando em treinamentos mais rápidos, mas com previsões menos precisas, fazendo com que o treinamento oscile na convergência de uma solução otimizada.

Logo após, finalizando a seção, na Tabela 9 são mostrados os melhores valores encontrados para cada hiperparâmetro – o *best tuning* – de cada modelo.

Tabela 8: Espaço de busca dos hiperparâmetros de cada algoritmo utilizado na avaliação experimental.

Algoritmo	Valores dos hiperparâmetros
RF	n_estimators: {5, 10, 100} max_features: {sqrt, log2} max_depth: {5, 10, 15} max_leaf_nodes: {3, 9, 15}
LR	C: {0.1, 1, 100} max_iter: {100, 1000, 100000} multi_class: {auto, ovr, multinomial}
SVM	kernel: {linear} C: {0.1, 1, 100} max_iter: {100, 1000, 50000}
NB	var_smoothing: {numpy.logspace(0,-9, num=100)}
AB	n_estimators: {1, 10, 100} algorithm: {SAMME, SAMME.R}
GB	learning_rate: {0.01, 0.05, 0.1, 0.2, 0.5} max_depth: {3, 9, 15} n_estimators: {1, 10, 100}
MLP	max_iter: {100, 1000, 10000} hidden_layer_sizes: {(150, 100, 50), (40, 80, 120), (100, 150, 30), (50, 50, 50)} activation: {tanh, relu, logistic} solver: {sgd, adam, lbfgs} learning_rate: {constant, adaptive}

Tabela 9: O *best tuning* de cada modelo.

Algoritmo	Configuração
RF	n_estimators: 100 max_features: sqrt
LR	C: 1.0 max_iter: 100 multi_class: auto
SVM	C: 1.0 max_iter: 1000
NB	var_smoothing: 5.336699231206313e-07
AB	n_estimators: 50 algorithm: SAMME.R
GB	learning_rate: 0.1 max_depth: 3 n_estimators: 100
MLP	max_iter: 10000 hidden_layer_sizes: (150, 100, 50) activation: tanh solver: lbfgs learning_rate: adaptive

5 RESULTADOS

Neste capítulo são evidenciados e discutidos os diferentes resultados obtidos, bem como analisadas as métricas de cada modelo treinado, objetivando a escolha e implantação (*deploy*) de um modelo que potencialmente resolva o problema de pesquisa aplicada em questão. Dá-se início à última etapa do processo: a extração do conhecimento adquirido nos padrões identificados.

5.1 Avaliação empírica

A Tabela 10 apresenta, para cada algoritmo de classificação, a qualidade da predição dos modelos sobre os dados de teste, considerando cada configuração de treinamento. As configurações de treinamento combinam dados desbalanceados (Desbal), balanceados (Bal), hiperparâmetros padrão (*def*) e hiperparâmetros ajustados (*tuned*). A métrica de qualidade apresentada é a média das μF_1 obtidas, considerando todos os *seeds*.

Os melhores resultados para cada algoritmo estão sublinhados enquanto os melhores de cada configuração de treinamento estão destacados em **negrito**. A tabela também apresenta a média dos resultados para cada algoritmo na última coluna, que foi utilizada para ordenação decrescente. Por fim, na última linha, são apresentadas as médias por conjunto de dados e configuração de hiperparâmetros.

Tabela 10: Desempenho sobre os dados de teste de todos os modelos treinados considerando a média ponderada da F1.

Algoritmo	Conjunto de dados de treinamento (hiperparâmetros)				Média
	Desbal (<i>def</i>)	Desbal (<i>tuned</i>)	Bal (<i>def</i>)	Bal (<i>tuned</i>)	
Random Forest	0,6797	0,7002	0,6807	0,6981	0,6897
Gradient Boosting	0,6811	<u>0,6837</u>	0,6740	0,6799	0,6797
Logistic Regression	0,6282	<u>0,6592</u>	0,5851	<u>0,6592</u>	0,6329
Adaboost	0,5520	<u>0,6687</u>	0,5142	<u>0,6687</u>	0,6009
Naive Bayes	0,6058	0,5784	0,5580	0,5784	0,5802
SVM	<u>0,4385</u>	0,4170	0,3384	0,4111	0,4013
MLP	0,3144	0,4001	0,3305	<u>0,4383</u>	0,3708
Média	0,5571	0,5868	0,5258	<u>0,5905</u>	0,5651

Analisando estes resultados, percebe-se que o algoritmo *Random Forest* obteve me-

lhor desempenho em três das quatro configurações dos modelos, perdendo para o *Gradient Boosting* apenas quando utilizada a distribuição original dos dados e configuração padrão de hiperparâmetros. O melhor resultado global, com $\mu F_1 = 0,7002$, foi atingido quando o algoritmo *Random Forest* foi treinado sobre os dados desbalanceados e com os hiperparâmetros ajustados. *Logistic Regression* e *Adaboost* mostraram-se invariantes ao balanceamento dos dados quando os hiperparâmetros foram devidamente ajustados, atingindo a mesma μF_1 igual a 0,6592 e 0,6687, respectivamente. Considerando apenas os resultados de modelos sintonizados (*tuned*), quatro dos sete algoritmos obtiveram melhores resultados quando treinados sobre os dados desbalanceados. Por fim, a melhor média dos resultados ($\mu F_1 = 0,5905$) foi atingida utilizando o SMOTE como estratégia de balanceamento dos dados e hiperparâmetros ajustados.

A Figura 11 sintetiza os dados da Tabela 10. Pode-se verificar pela área formada pelas curvas apresentadas que, independente da configuração de treinamento, o algoritmo *Random Forest* obteve o melhor desempenho médio, seguido pelo algoritmo *Gradient Boosting*. Este, além de similar no desempenho, apresentou também regularidade em relação às diferentes configurações de experimentos. Em contrapartida, os algoritmos SVM e MLP obtiveram os piores desempenhos. Vale ressaltar que não era objetivo deste trabalho encontrar o melhor *kernel* ou a melhor arquitetura de rede neural e, portanto, diferentes configurações destes algoritmos não experimentadas neste trabalho poderiam obter resultados consideravelmente melhores.

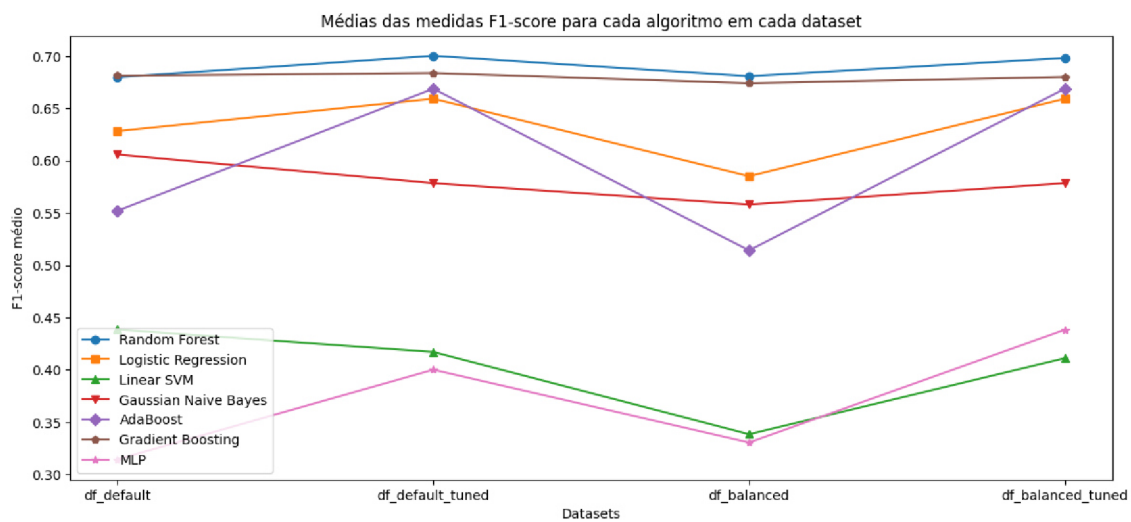


Figura 11: Comparação de desempenho dos algoritmos em diferentes configurações de treinamento.

Visando aprofundar a avaliação individual dos modelos através de análises gráficas, para cada conjunto de resultados obtidos pela variação das *seeds*, foram plotados *boxplots* para evidenciar, de forma comparativa, a distribuição das μF_1 obtidas.

A Figura 12 considera o desempenho de cada modelo treinado sobre os dados desbalanceados com hiperparâmetros *default*. Pode-se verificar que enquanto os algoritmos

Random Forest e *Gradient Boosting* apresentam desempenho positivos similares, ou seja, pouca variância sobre medianas relativamente altas, os algoritmos SVM e MLP apresentam os piores resultados, com uma grande variação sobre o centro das distribuições.

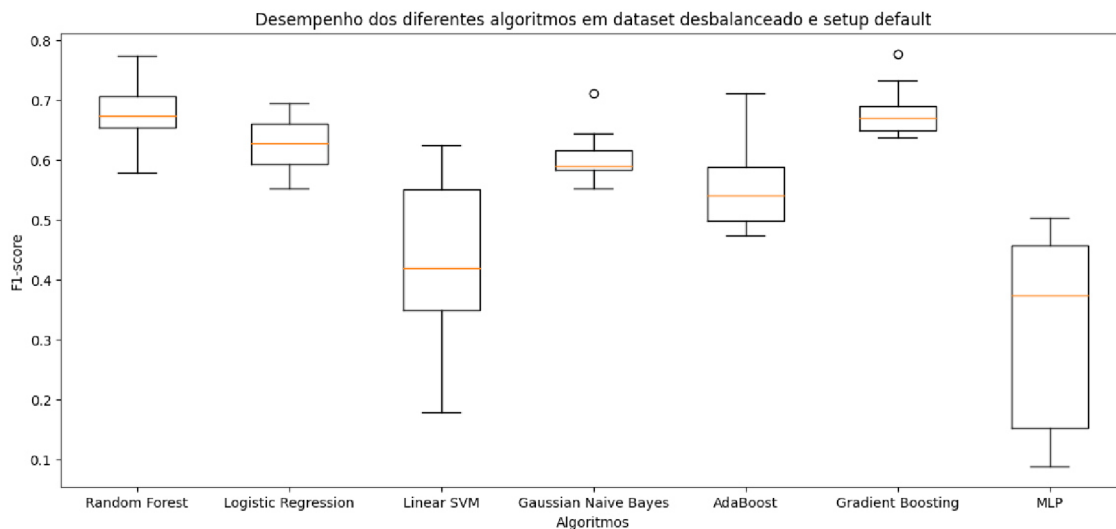


Figura 12: Distribuição da μF_1 considerando dados desbalanceados sem ajuste de hiperparâmetros.

Já na Figura 13, observa-se que o ajuste de hiperparâmetros alterou o desempenho dos algoritmos aplicados sobre os mesmos dados. A distribuição do desempenho dos algoritmos *Random Forest* e *Gradient Boosting* – até então os dois algoritmos com melhor desempenho – têm sua variação levemente incrementada. As distribuições dos algoritmos SVM e MLP tiveram uma melhora substancial, mas permaneceram com desempenho inferior aos demais. *Adaboost* também se mostrou sensível positivamente ao ajuste dos hiperparâmetros, sendo comparável aos melhores: *Random Forest*, *Gradient Boosting* e *Logistic Regression*.

Com os dados balanceados, a Figura 14 mostra o comparativo dos algoritmos. É possível depreender que o resultado da imputação de instâncias sintéticas nas classes minoritárias, sem o ajuste dos hiperparâmetros, resultou em desempenhos similares aos dados desbalanceados e igualmente sem ajuste dos hiperparâmetros. Se faz necessário mais – e diferentes – análises, mas, até o momento, pode-se inferir a hipótese de que o desbalanceamento das classes, considerando o conjunto de dados utilizado na presente pesquisa, não modificou o desempenho dos modelos.

Por último, na Figura 15, pode-se verificar que o desempenho dos algoritmos foi superior quando os hiperparâmetros são ajustados, ou seja, quando comparados aos resultados da Figura 14. Os algoritmos com desempenho mais influenciado pelo ajuste foram SVM, MLP e *Adaboost*. Este último, novamente, obteve desempenho similar ao dos melhores algoritmos.

Após as análises gráficas dos diferentes algoritmos aplicados às diferentes

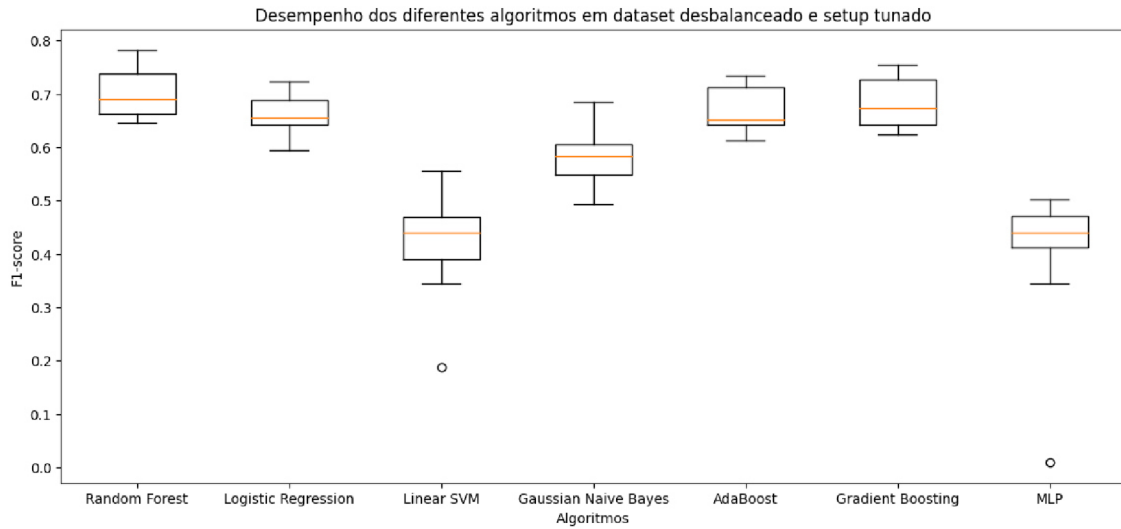


Figura 13: Distribuição da μF_1 considerando dados desbalanceados e hiperparâmetros ajustados.

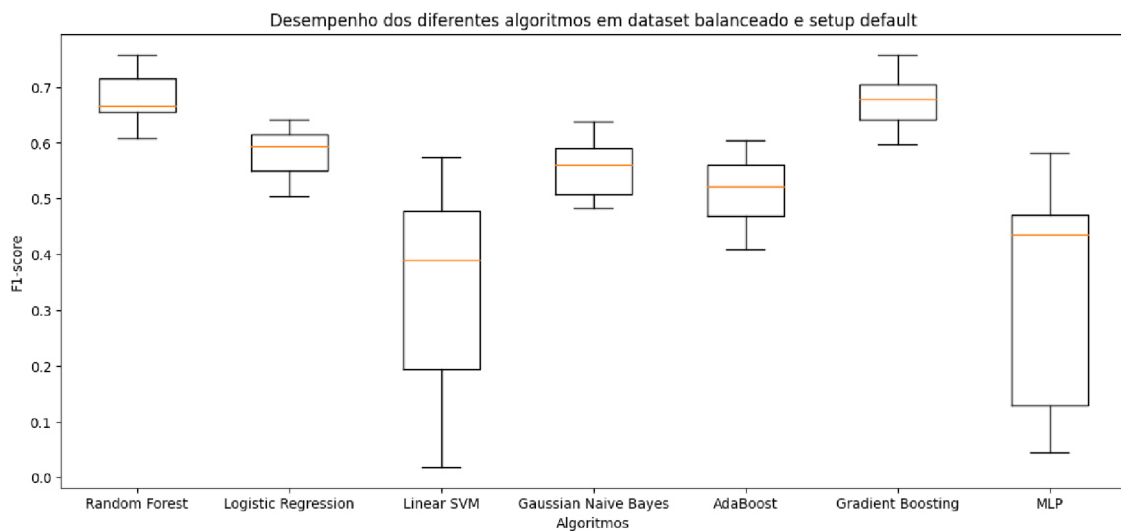


Figura 14: Distribuição da μF_1 considerando dados balanceados sem ajuste de hiperparâmetros.

configurações de treinamento, é necessária a exploração dos desempenhos compilados, de forma a entender o quanto as diferentes configurações relacionadas ao balanceamento e ao ajuste de hiperparâmetros influenciam no desempenho dos modelos. A Figura 16 demonstra tais diferenças, possibilitando concluir que o balanceamento não influenciou positivamente qualquer modelo, sendo responsável por piorar a qualidade dos modelos sem ajuste de hiperparâmetros.

Finalizando a presente seção, pode-se concluir que as avaliações individuais dos algoritmos, aplicados às diferentes configurações de treinamento dos modelos corroboram com a avaliação inicial desta mesma seção, na qual os modelos construídos utilizando o algoritmo *Random Forest* obtiveram as maiores médias de todas as μF_1 comparadas,

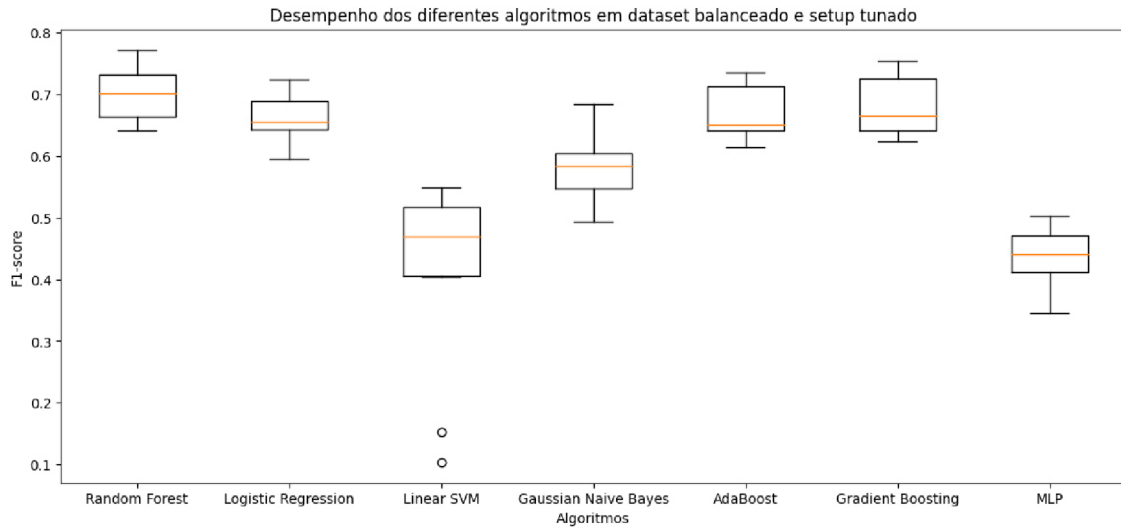


Figura 15: Distribuição da μF_1 considerando dados balanceados e hiperparâmetros ajustados.

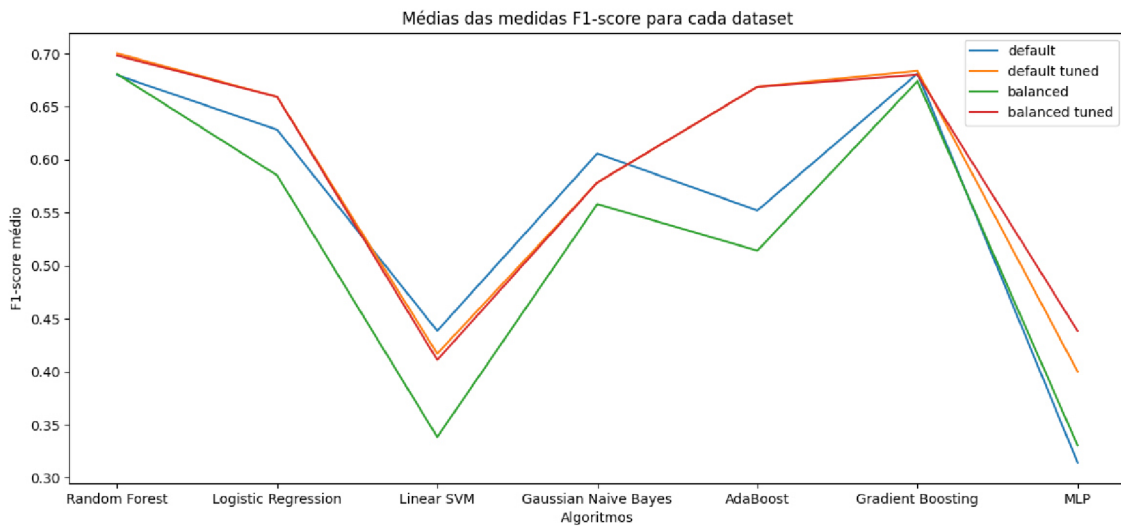


Figura 16: Métricas F1 de cada modelo em cada configuração dos dados.

seguidos pelo algoritmo *Gradient Boosting*.

5.2 Análise estatística

A análise experimental sobre o desempenho de um método proposto é tarefa crucial – e necessária – a ser realizada em uma pesquisa. Por isso, se faz importante uma análise estatística sobre os resultados obtidos pelos modelos em diferentes situações. Estas situações, no presente trabalho, são emuladas justamente pelos subconjuntos de dados distintos gerados pelas diferentes *seeds*.

Em se tratando de inferências estatísticas, o primeiro passo consiste em analisar os dados que se têm à disposição. No presente caso, os dados são os valores da métrica

F1 para cada *seed*, em cada modelo gerado por um diferente algoritmo, com diferentes hiperparâmetros configurados e em diferentes dados – balanceados e desbalanceados. Os valores de F1 não compõem uma distribuição normal, então, os testes estatísticos mais adequados são os não-paramétricos. Testes não-paramétricos podem ser empregados satisfatoriamente para comparar classificadores genéricos em vários conjuntos de dados considerando qualquer medida de desempenho [27].

Particularmente, o teste estatístico escolhido neste trabalho foi o *Aligned-Friedman* [35] e, após este teste, é apresentado o cálculo dos *p-values* ajustados no teste *post hoc* de *Holm* [36], de forma ordenada, evidenciando o modelo com melhor desempenho estatístico – e sugerido neste trabalho.

O teste *Aligned-Friedman* é um método estatístico utilizado para analisar dados coletados de várias condições experimentais. Este teste é uma extensão do teste de Friedman – que é uma versão não-paramétrica do teste ANOVA [79] – que lida com uma questão comum em amostras com medidas repetidas, indicado para comparações de algoritmos de medidas repetidas.

Basicamente, este teste é subdividido internamente em algumas etapas, as quais: (1) organização das amostras em uma matriz, alinhando-as de acordo com as coletas; (2) para cada par de amostras, são calculadas as diferenças; (3) tais diferenças são submetidas ao teste de Friedman padrão, para determinar se há diferenças estatisticamente significativas entre as amostras; e, por último, (4) o *p-value* obtivo a partir do teste de Friedman é ajustado.

O teste de *Holm* é considerado um teste *post hoc*, ou seja, um teste estatístico adicional, realizado após a aplicação – e análise – de um teste estatístico principal, como o *Aligned-Friedman*, aplicado neste trabalho. O objetivo de um teste *post hoc* é apresentar comparações múltiplas, ou seja, apresentar comparações adicionais entre as amostras.

Através do teste de *Holm*, é possível identificar quais amostras apresentam diferenças significativas, comparando as amostras entre si – elegendo uma delas como sendo a variável-controle – e controlando o erro global.

Neste trabalho, os testes foram realizados através da plataforma STAC [70], disponível na Web¹, com um nível de significância estabelecido em 0,05 (ou 5%) para limite de rejeição da hipótese nula. Inicialmente, os testes foram realizados com cada um dos quatro conjuntos, conforme descrito a seguir:

- Conjunto 1 – dados desbalanceados e hiperparâmetros em modo *default*;
- Conjunto 2 – dados desbalanceados e hiperparâmetros ajustados com o *best tuning*;
- Conjunto 3 – dados balanceados e hiperparâmetros em modo *default*;
- Conjunto 4 – dados balanceados e hiperparâmetros ajustados com o *best tuning*.

¹<https://tec.citius.usc.es/stac/>

Sobre o conjunto 1, o *p-value* não-corrigido resultou em um valor de 0.005941, ou seja, um valor de aproximadamente 0,6%. Como o nível de significância estabelecido para este teste foi de 5%, pode-se afirmar que a hipótese nula – de que todos os algoritmos obtiveram o mesmo desempenho, em se tratando da métrica escolhida, a F1 – foi rejeitada. Ou seja, o teste *Aligned-Friedman* pode afirmar que, no conjunto 1, pelo menos um dos modelos é estatisticamente diferente dos demais.

Na Tabela 11 são apresentados os resultados do teste *Aligned-Friedman* com o teste *post hoc* de *Holm*, em formato de *ranking*, com os *p-values* corrigidos, os valores de estatística calculados, e o resultado do teste de hipóteses sobre o conjunto 1.

Tabela 11: *Ranking* após os testes estatísticos sobre o conjunto 1.

<i>Ranking</i>	<i>Algoritmo</i>	<i>p-value corrigido</i>	<i>Estatística</i>	<i>Hipótese nula</i>
6.50	Random Forest	N/A	N/A	N/A
6.30	Gradient Boosting	0,85513	0,18257	Não rejeitada
4.50	Logistic Regression	0,20367	1,82574	Não rejeitada
4.00	Naive Bayes	0,08992	2,28218	Não rejeitada
3.30	Adaboost	0,01744	2,92119	Rejeitada
2.05	SVM	0,00029	4,06228	Rejeitada
1.35	MLP	0,00002	4,70129	Rejeitada

Nos resultados apresentados após os testes estatísticos no decorrer desta seção, as colunas de rótulo *ranking* evidenciam o algoritmo com o melhor desempenho – quanto mais alto o valor apresentado nesta coluna, melhor o desempenho do algoritmo sobre os dados específicos; as colunas *algoritmo* descrevem o algoritmo a ser comparado com o método de controle; as colunas *p-value corrigido* indicam a probabilidade de, nos testes estatísticos, se obter resultados falso-positivos²; as colunas *estatística* demonstram os valores de estatística de teste, que são utilizados para avaliar as diferenças entre o algoritmo comparado e o método de controle; por último, as colunas *hipótese nula* indicam se, na comparação de cada algoritmo com o método de controle, se a hipótese nula – de que os algoritmos são estatisticamente iguais – é rejeitada ou não rejeitada.

Nos testes estatísticos realizados sobre o conjunto 1 (dados desbalanceados e hiperparâmetros em modo *default*), apresentados na Tabela 11, o algoritmo *Random Forest* obteve o melhor desempenho e foi escolhido como método de controle, ou seja, como ponto de referência a ser utilizado para comparar o comportamento de todos os demais algoritmos em relação a ele. Pode-se afirmar que existiram diferenças significativas quando comparado com os algoritmos MLP, *Adaboost* e SVM. No comparativo do *Random Forest* com os algoritmos *Gradient Boosting*, *Logistic Regression* e *Naive Bayes*, não foi possível detectar diferenças.

Dando prosseguimento à análise, é preciso realizar os mesmos testes sobre os demais

²No campo da estatística, de forma simples, esta é a definição de *erro tipo I*.

conjuntos. Então, os resultados dos testes de *Friedman* foram compilados para um entendimento facilitado, conforme segue abaixo:

- Conjunto 1 – *p-value* igual a 0,5941%;
- Conjunto 2 – *p-value* igual a 0,0158%;
- Conjunto 3 – *p-value* igual a 1,2309%;
- Conjunto 4 – *p-value* igual a 0,2495%;

Pode-se observar, através dos testes *Aligned-Friedman* sobre os quatro conjuntos que, em todos eles, o *p-value* ficou abaixo do nível de significância estabelecido de 5%, ou seja, houve rejeição da hipótese nula em todos os conjuntos – o que, em outras palavras, é o mesmo que afirmar que, em cada um dos conjuntos de dados, pelo menos um dos algoritmos é estatisticamente diferente dos demais.

As Tabelas 12, 13 e 14 demonstram a aplicação dos testes *Aligned-Friedman* e *Holm* dispostos em cada um dos demais conjuntos, evidenciando o *ranking* de desempenho dos algoritmos em cada conjunto.

Tabela 12: *Ranking* após os testes estatísticos sobre o conjunto 2.

Ranking	Algoritmo	<i>p-value</i> corrigido	Estatística	Hipótese nula
6.70	Random Forest	N/A	N/A	N/A
5.70	Gradient Boosting	0,47067	0,91287	Não rejeitada
4.90	Adaboost	0,30104	1,64317	Não rejeitada
4.70	Logistic Regression	0,27156	1,82574	Não rejeitada
3.00	Naive Bayes	0,00366	3,37762	Rejeitada
1.65	SVM	0,00002	4,61000	Rejeitada
1.35	MLP	0,00001	4,88386	Rejeitada

Tabela 13: *Ranking* após os testes estatísticos sobre o conjunto 3.

Ranking	Algoritmo	<i>p-value</i> corrigido	Estatística	Hipótese nula
6.60	Random Forest	N/A	N/A	N/A
6.40	Gradient Boosting	0,85513	0,18257	Não rejeitada
4.40	Logistic Regression	0,13383	2,00832	Não rejeitada
3.90	Naive Bayes	0,05484	2,46475	Não rejeitada
3.10	Adaboost	0,00699	3,19505	Rejeitada
1.90	SVM	0,00011	4,29049	Rejeitada
1.70	MLP	0,00005	4,47307	Rejeitada

Em todos os conjuntos e configurações dos dados testados, o algoritmo *Random Forest* permaneceu no topo das análises e foi considerado o método de controle nas comparações estatísticas *post hoc* entre todos os demais algoritmos – um indício importante da confiabilidade nos seus resultados. Outro ponto que pode-se depreender é quanto à equivalência

Tabela 14: *Ranking* após os testes estatísticos sobre o conjunto 4.

Ranking	Algoritmo	<i>p-value</i> corrigido	Estatística	Hipótese nula
6.80	Random Forest	N/A	N/A	N/A
5.60	Gradient Boosting	0,54664	1,09545	Não rejeitada
4.90	Adaboost	0,24851	1,73445	Não rejeitada
4.70	Logistic Regression	0,22094	1,91703	Não rejeitada
2.90	Naive Bayes	0,00185	3,56020	Rejeitada
1.70	SVM	0,00002	4,65564	Rejeitada
1.40	MLP	0,00001	4,92950	Rejeitada

das performances dos algoritmos face aos diferentes conjuntos: são claras as diferenças estatísticas entre os diferentes algoritmos, mantendo-se permanentes na não rejeição da hipótese nula os algoritmos *Random Forest*, *Gradient Boosting* e *Logistic Regression*, ao passo que os algoritmos SVM e MLP mantiveram-se igualmente permanentes, mas no sentido oposto, ou seja, na rejeição da hipótese nula frente ao método de controle.

Faz-se importante notar que os resultados do *ranking* dos testes estatísticos realizados neste trabalho, utilizando as amostras de tamanho igual a 10 (sendo cada *seed* uma representante de cada amostra), traduzem fielmente o *ranking* retornado na avaliação empírica que foi realizada neste trabalho (Seção 5.1), mantendo exatamente as mesmas posições dos algoritmos para cada conjunto de dados.

Após a análise dos *rankings* – levando em consideração a métrica escolhida, a F1 – dos algoritmos nos quatro diferentes conjuntos e configurações dos dados, pode-se afirmar que os resultados não foram baseados em uma possível variabilidade dos dados, podendo o algoritmo *Random Forest*, num viés estatístico, ser considerado superior aos algoritmos que tiveram, na comparação, a hipótese nula rejeitada em cada conjunto de dados. De igual forma, o algoritmo *Random Forest* também pode ser considerado – no mínimo – estatisticamente equivalente aos algoritmos os quais, na comparação, a hipótese nula não foi rejeitada (ainda que, na amostra considerada nesta dissertação, este método tenha sido considerado superior aos demais).

Por fim, a Tabela 15 demonstra a configuração dos melhores valores para os hiperparâmetros do referido modelo.

Tabela 15: Melhor configuração de hiperparâmetros do melhor modelo encontrado.

Hiperparâmetro	Valor
n_estimators	100
max_features	sqrt

5.3 Melhor modelo

Após o desenvolvimento dos modelos³ com capacidades preditivas suficientes para resolver o problema de pesquisa dessa dissertação, esta seção apresenta um estudo de caso baseado do melhor dos modelos, considerando tanto a avaliação empírica (Seção 5.1) quanto a análise estatística (Seção 5.2) realizadas.

Neste trabalho, foram calculadas diferentes métricas, em 10 diferentes *seeds*, representando uma amostra de tamanho igual a 10 da população de cada configuração dos dados. Com os dados disponíveis na análise empírica, pode-se depreender que o algoritmo com melhor desempenho na resolução do presente problema de pesquisa é o *Random Forest*. Com relação à análise estatística, há que se considerar que o *Random Forest* foi o algoritmo melhor posicionado no *ranking* do teste de *Friedman*, servindo como método de controle para o teste *post hoc* de *Holm*, o que o caracteriza como uma opção recomendável⁴.

O algoritmo a ser considerado para a busca do *melhor modelo* será o *Random Forest*. Deste algoritmo, sabe-se a melhor configuração de hiperparâmetros (*best tuning*) possível para que sua capacidade de predição, bem como sua capacidade de generalização, sejam otimizadas.

Para encontrar o melhor modelo dentre todos os modelos desenvolvidos utilizando o algoritmo *Random Forest*, é necessário encontrar a *seed* que possui o valor mais alto de μF_1 . Será esta *seed*, representando uma amostragem de todas as demais *seeds* utilizadas, que será detalhada e considerada como o melhor modelo – sugerido para utilização industrial *a posteriori* neste trabalho. Nesse sentido, a *seed* com valor mais alto de μF_1 , para este algoritmo, neste trabalho, foi a *seed* de número 100000. A Tabela 16 apresenta a matriz de confusão do referido modelo, confrontando sua capacidade preditiva com os dados de teste disponíveis.

Tabela 16: Matriz de confusão do algoritmo *Random Forest*, sobre dados balanceados e hiperparâmetros ajustados através do seu *best tuning*, na *seed* 100000.

	[0-10]	(10-20]	(20-30]	(30-inf]
[0-10]	53	4	3	1
(10-20]	7	16	6	0
(20-30]	1	2	3	0
(30-inf]	0	0	2	0

Através da análise da matriz de confusão, pode-se calcular a acurácia do modelo, igual

³Por *desenvolvimento dos modelos*, entende-se as etapas de pré-processamento dos dados, treinamento dos modelos, validação e teste.

⁴Estatisticamente, não se pode afirmar que o algoritmo *Random Forest* é superior aos demais algoritmos, pois há que se considerar que uma amostra pode apresentar um resultado, enquanto outra amostra pode apresentar outro resultado – a isso, chama-se *variabilidade dos dados*. Entretanto, independente das amostras, pode-se afirmar que, estatisticamente, o algoritmo *Random Forest* tem a performance tão boa quanto os demais algoritmos os quais, na comparação, não rejeitaram a hipótese nula.

a 73%. Pode-se depreender que sua capacidade preditiva seja considerada suficiente em processos produtivos, ou seja, para o uso industrial da madeira. Entretanto, para que este objetivo seja atingido, é preciso aprofundar a análise, verificando as métricas para cada classe, conforme evidencia a Tabela 17.

Tabela 17: Métricas do algoritmo *Random Forest* para cada classe, sobre dados balanceados e hiperparâmetros ajustados através do seu *best tuning*, na *seed* 100000.

Métrica	[0-10]	(10-20]	(20-30]	(30-inf]
Precisão	87%	73%	21%	0%
Revocação	87%	55%	50%	0%
F1	87%	63%	30%	0%
Suporte	61	29	6	2

Verifica-se, através da matriz de confusão gerada nos parâmetros supracitados, que mesmo após as diferentes abordagens utilizadas, o melhor dos modelos desenvolvido deste trabalho foi capaz de prever corretamente apenas metade das instâncias de teste da classe (20-30], enquanto da classe (30-inf], o modelo não foi capaz de prever corretamente nenhuma instância.

No presente problema, pode-se verificar que todas as métricas – precisão, revocação e F1 – possuem valores aceitáveis na classe majoritária, e decrescem de maneira inversamente proporcional ao suporte (número de instâncias) em direção à classe minoritária.

Os valores das métricas chegam a ser zero na classe (30-inf], ou seja, não há qualquer predição correta desta classe. Na classe (20-30], com suporte igual a 6, os valores são levemente maiores, mas, ainda assim, não suficientes para sugestão de uso industrial *a posteriori*. Por outro lado, pode-se afirmar que as mesmas métricas aplicadas sobre as classes originalmente majoritárias – [0-10] e (10-20] – possuem valores considerados adequados: a menor precisão do modelo, neste caso, foi de 73% para a classe (10-20], com 29 instâncias como suporte, enquanto os maiores valores das métricas foram obtidos na classe [0-10], com 61 instâncias de suporte, com valores próximos de 90%.

Evidencia-se, então, através da análise das métricas por classe que, apesar de o modelo possuir baixa capacidade preditiva sobre as classes minoritárias, o mesmo possui alta capacidade preditiva sobre as classes majoritárias. Portanto, para que seja possível analisar a capacidade preditiva de forma ainda mais aprofundada e entender o real poder de classificação do modelo proposto, é preciso analisar os casos de falha do modelo, ou seja, buscar entender o motivo pelo qual o modelo optou por uma determinada classe – errada – em detrimento da classe correta.

5.3.1 Análise dos casos de falha

Avaliar um modelo de predição, quando se tem por objetivo uma classificação, pode não ser uma tarefa trivial em problemas do mundo real. Variáveis que, em situações

diversas, seriam absolutamente ignoradas, podem passar a ter tamanha importância que, em função delas, o modelo pode ter até mesmo sua utilização condenada.

No caso em questão, sabe-se que, apesar de a acurácia do modelo ser considerada aceitável – e suficiente para que o modelo pudesse ser sugerido – os dados originalmente desbalanceados comprometem a utilização desta métrica. Já as métricas ponderadas (precisão, revocação e F1), utilizadas igualmente nesta dissertação, têm seus valores diretamente proporcionais ao número de instâncias como suporte de cada classe no conjunto de teste, ou seja, quanto mais instâncias uma classe possui no conjunto de teste, maior é a capacidade preditiva do modelo em questão.

Até este ponto, é impossível conhecer a real capacidade preditiva, já que o modelo tem excelente desempenho em classes majoritárias, e fraco desempenho em classes minoritárias. O modelo continuará classificando corretamente se houver mais instâncias? Ou o modelo continuará classificando erroneamente se houver menos instâncias? Como será o comportamento do modelo frente a problemas do mundo real? Para responder a estas questões, faz-se imprescindível uma análise detalhada justamente das instâncias nas quais o modelo foi incapaz de predizê-las corretamente – o que tem por nome a *análise dos casos de falha* do modelo.

Das 98 instâncias presentes nos dados de teste, o modelo não foi capaz de prever corretamente 27 delas. Antes de aprofundar a análise, é preciso conhecer a importância de cada atributo, considerado pelo modelo, para realizar uma determinada predição. A Figura 17 evidencia os dez atributos mais importantes⁵ para o modelo desenvolvido.

Através da análise da figura, pode-se verificar que, de maneira decrescente, estão dispostos os dez atributos mais importantes para o modelo cuja análise está sendo aprofundada. Verifica-se que o atributo mais importante para a predição da variabilidade do teor de umidade na madeira é a quantidade de madeira que foi depositada na pilha de toras (atributo *TOT_ENTR*), seguido pelo tempo no qual as toras permaneceram armazenadas na pilha (atributo *QTD_DIAS*). Ambos atributos são provenientes dos *dados operacionais* e possuem relação direta com o decréscimo de peso, pois sabe-se que a variação do peso da madeira armazenada ao ar livre (secagem natural) tem relação direta com a variação do teor de umidade.

Logo após, com uma certa diferença de importância, estão mais três atributos: a umidade relativa média do período (atributo *UR_MED*), a temperatura máxima atingida no período de armazenamento (atributo *TEMP_MAX*) e a precipitação, ou seja, a quantidade de chuva observada durante o período (atributo *PRECIPITACAO*). Tais atributos são provenientes dos *dados meteorológicos*, demonstrando, assim, que esta categoria de dados, apesar de ser muito importante – e imprescindível – para a qualidade das predições do modelo, não compõe o primeiro escalão de atributos prioritários.

⁵As importâncias dos atributos são fornecidas através de um método do classificador, sendo calculadas através da diminuição de impurezas dentro de cada árvore.

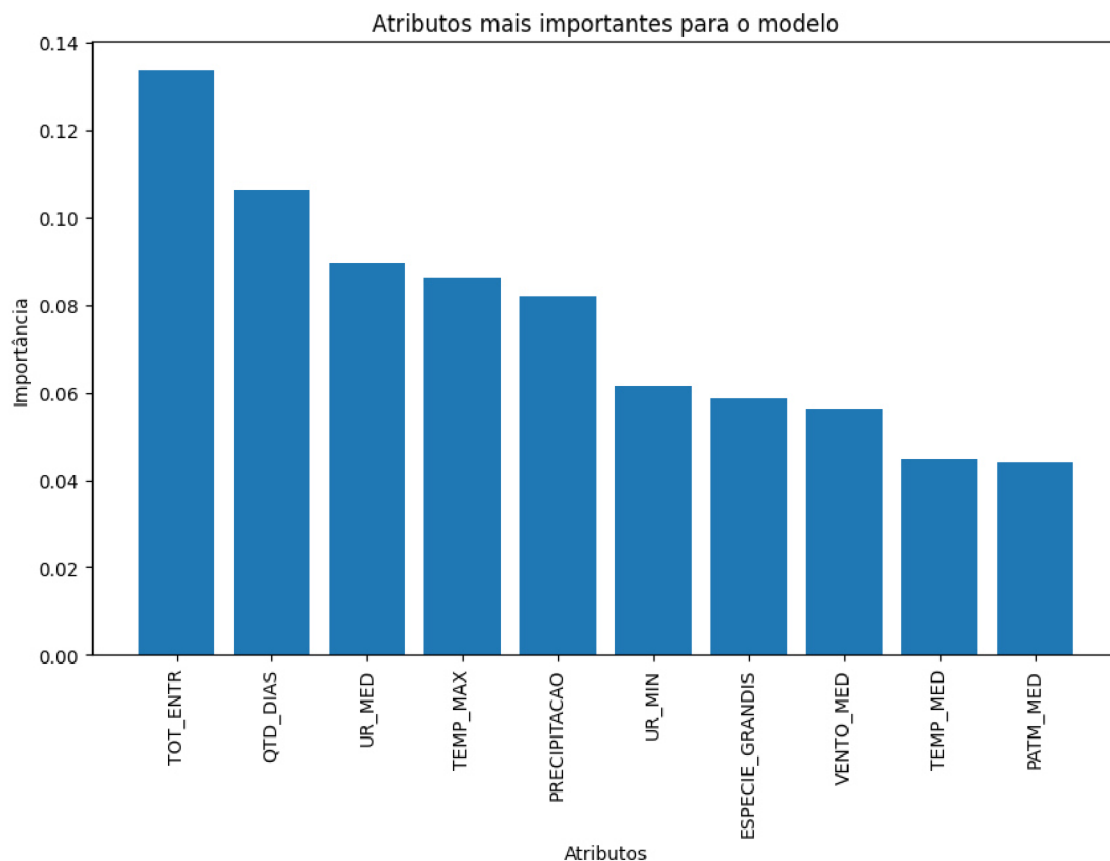


Figura 17: Importância de cada atributo para o modelo analisado, considerando os dez atributos mais importantes.

Compondo o terceiro escalão, estão mais três atributos: a umidade relativa mínima registrada no período (atributo *UR_MIN*), a espécie da madeira armazenada na pilha – quando esta é do tipo *GRANDIS*, e o velocidade média do vento registrado no período (atributo *VENTO_MED*). Destes três atributos, é importante pontuar que a espécie *GRANDIS* é a única espécie dentre os dez atributos mais importantes considerados pelo modelo – potencialmente, é a espécie que é mais suscetível à variação do teor de umidade em função do tempo de armazenamento.

As últimas duas variáveis que compõem as dez mais importantes são a temperatura média do período (atributo *TEMP_MED*) e a pressão atmosférica média registrada no mesmo período (atributo *PATM_MED*) que, ainda que sejam consideradas muito importantes para o modelo, sua importância não chega à metade da média da importância dos atributos do primeiro escalão (*TOT_ENTR* e *QTD_DIAS*).

Conhecida a importância de cada atributo considerado nas predições, é possível analisar de maneira crítica os casos de falha do referido modelo. A Tabela 18 mostra todas as 27 instâncias nas quais o modelo foi incapaz de prever corretamente seu intervalo de variação do teor de umidade, considerando apenas os dez atributos mais importantes para as suas predições.

Na tabela que apresenta os casos de falha, cada linha corresponde uma observação, ou uma instância na qual o modelo foi incapaz de prever corretamente o intervalo de perda do peso em função do teor de umidade, sendo referenciada através da primeira coluna – o índice – representando o número da instância no conjunto de dados original.

Pode-se notar algumas discrepâncias nos dados originais. Nas instâncias de índices 7, 20, 259, 274, 523 e 637, o total de entrada de madeira (em Kg) em cada pilha foi igual ou inferior a duas cargas de um caminhão florestal – com peso médio de 32.000 Kg – durante todo o período observado. Em teoria, as pilhas são locais de armazenamento que recebem grandes quantidades de toras de madeira, e a média das entradas de madeira nas pilhas, levando em consideração o conjunto de dados original, é de 216.740 Kg⁶.

De igual forma, pode-se notar discrepâncias também nas instâncias de índices 232, 259, 260 e 754 em relação à quantidade de dias em que as toras de madeira permaneceram armazenadas nas pilhas. Tais instâncias mantiveram suas cargas armazenadas por períodos variando entre 1 e 9 dias, apenas – o que inviabilizaria operacionalmente a criação de uma nova pilha para armazenar carga somente por poucos dias. Já as instâncias de índices 11, 22, 100, 417, 528, 596 e 625 percorreram o caminho oposto, mantendo suas cargas armazenadas por, no mínimo, mais de 152 dias – equivalente a praticamente cinco meses de armazenamento – enquanto a média do tempo armazenado no conjunto de dados original é de 47 dias⁷.

Outra possível discrepância é possível de ser depreendida se for considerada a relação do total de entradas com a quantidade de dias de armazenamento. Nas instâncias de índices 358 e 444, o total de entradas foi de 471.780 Kg e 300.240 Kg, respectivamente, enquanto o total de dias de armazenamento foi, igualmente de forma respectiva, de 121 e 126. Para estas instâncias, o intervalo de perda real de peso foi de [0-10]. Considerando a quantidade de dias de armazenamento, pode-se concluir que a média das toras de madeira armazenadas nestas pilhas já havia sido cortada há mais tempo.

Nas 9 instâncias remanescentes, não foi possível identificar possíveis discrepâncias operacionais, podendo tais instâncias serem consideradas como *falhas efetivas* do modelo.

Ao analisar os casos de falha de um modelo, é possível extrair observações importantes sobre a verdadeira capacidade preditiva do mesmo, possibilitando a identificação de discrepâncias operacionais presentes nos dados, além de uma melhor compreensão das características de funcionamento do modelo e, também, de possíveis padrões dos dados que o modelo possa ter negligenciado. Tais premissas permitem não só que o modelo seja melhor avaliado, mas também acarretam em decisões operacionais relevantes – como aprimorar a coleta de dados ou, até mesmo, fornecer informações às empresas geradoras dos dados para que observem seus processos operacionais.

⁶O valor da média do total de entradas leva em consideração, também, os valores discrepantes observados durante a análise dos casos de falha.

⁷De igual modo, o valor da média da quantidade de dias de armazenamento leva em consideração, também, os valores discrepantes observados durante a análise dos casos de falha.

5.4 Ética na utilização do modelo

Em se tratando de inovação – mais especificamente na área da tecnologia – o esperado é que os avanços obtidos precedam as diretrizes morais referentes à sua própria utilização. O princípio de moralidade não consiste em roteiros pré-definidos mas, sim, em uma bússola que continuamente aponta para o norte [15].

No caso específico da utilização do modelo sugerido nesta dissertação por organizações ligadas diretamente ao mercado da madeira, num primeiro olhar, levantam-se alguns questionamentos: a tecnologia desenvolvida neste trabalho pode ser empregada no mercado de extração ilegal de madeira? A solução contribui, de alguma forma, na diminuição do desmatamento?

No decorrer desta dissertação, foram desenvolvidos aproximadamente 280 modelos, utilizando diversos algoritmos e configurações de hiperparâmetros, com derivações do conjunto original de dados. De fato, não houve preocupação com a utilização da solução pelo mercado ilegal de madeira e poderia, portanto, a solução ser utilizada sem prejuízo.

Entretanto, uma parte do pressuposto utilizado para o treinamento dos modelos é o tipo de madeira (*pinus* e *eucalipto*) e a espécie. Os modelos treinados foram adaptados única e, tão somente, às possibilidades inseridas no conjunto de dados fornecido neste trabalho. Tanto o *pinus* quanto o *eucalipto* são consideradas árvores de florestas plantadas [41], ou seja, árvores que não são nativas e, portanto, sendo consideradas como *madeira de reflorestamento com foco em fins comerciais* – podendo sua extração ser realizada de forma legalizada.

Para que a tecnologia sugerida nesta dissertação possa ser utilizada pela indústria que se beneficia da extração ilegal de madeira, possivelmente seria necessária a criação de um novo modelo, utilizando novas *features*, com potencialmente diferentes ajustes nos hiperparâmetros e, até mesmo, na aderência dos algoritmos aos dados de treinamento, pois a característica do tipo e da espécie de madeira seriam diferentes.

Dessa forma, pode-se concluir que o modelo sugerido neste trabalho não pode ser transferível – sem que seja realizada uma análise com maior profundidade – para aplicações industriais baseadas em extração de madeira ilegal.

Tabela 18: Casos de falha do modelo.

Índice	Tot. Ent.	Qtd. dias	UR (Méd.)	Temp. Máx.	Precip.	UR (Min)	Espéc.	Vento (Méd.)	Temp. Méd.	Pressão (Méd.)	Classe real	Predição
7	29920	134	70.86	39.4	223.4	21	ELLIOTTI	3.19	22.81	1012.28	(20-30]	(10-20]
11	219060	171	77.88	34.4	504.6	21	DUNNI	3.17	15.78	1017.17	(30-inf]	(20-30]
20	34700	181	77.85	34.4	538.4	21	DUNNI	3.18	15.86	1016.90	[0-10]	(20-30]
22	181860	162	74.66	39.4	453.8	21	GRANDIS	3.30	19.43	1014.34	[0-10]	(20-30]
100	223260	152	77.74	34.4	461.2	21	DUNNI	3.14	16.21	1016.74	[0-10]	(20-30]
232	128480	1	75.17	23.1	3.2	45	GRANDIS	2.22	15.10	1024.18	(10-20]	[0-10]
259	60240	3	70.34	25.0	10.4	30	GRANDIS	3.21	15.60	1019.37	(20-30]	[0-10]
260	167960	9	69.32	28.9	19.2	30	GRANDIS	3.79	20.44	1011.08	(10-20]	[0-10]
274	45720	139	77.20	34.4	439.8	21	GRANDIS	3.18	16.40	1016.56	[0-10]	(30-inf]
327	133480	76	73.50	34.4	213.0	21	DUNNI	3.31	18.46	1015.36	(10-20]	(20-30]
358	471780	121	72.10	39.4	184.6	21	GRANDIS	3.29	20.79	1013.69	[0-10]	(10-20]
390	360940	135	76.77	34.4	444.0	21	DUNNI	3.23	16.66	1016.34	(10-20]	(20-30]
417	221400	229	76.35	39.4	605.8	21	DUNNI	3.20	17.47	1015.97	(10-20]	(20-30]
428	400720	50	71.16	32.5	98.0	24	GRANDIS	3.09	22.53	1012.46	(10-20]	[0-10]
444	300240	126	76.80	34.4	398.8	21	GRANDIS	3.22	17.08	1016.20	[0-10]	(10-20]
523	76340	425	76.06	39.4	1156.0	21	N/A	3.05	17.73	1015.24	[0-10]	(30-inf]
528	125960	170	77.75	34.4	524.6	21	GRANDIS	3.20	16.00	1016.80	[0-10]	(30-inf]
572	203680	53	71.54	34.4	90.4	21	GRANDIS	3.37	19.75	1014.00	(10-20]	[0-10]
591	88800	144	73.56	39.4	353.0	21	GRANDIS	3.36	20.1	1013.99	(30-inf]	(20-30]
596	167180	167	73.59	39.4	459.6	21	ELLIOTTI	3.32	20.25	1013.78	(10-20]	(20-30]
607	150520	118	76.89	34.4	391.6	21	GRANDIS	3.26	17.36	1015.86	(10-20]	(30-inf]
625	170220	201	75.97	39.4	528.6	21	GRANDIS	3.18	17.99	1015.63	(20-30]	(10-20]
637	48940	136	73.15	39.4	348.2	21	GRANDIS	3.35	19.95	1014.00	(10-20]	(20-30]
676	325160	77	70.56	39.4	113.0	21	GRANDIS	3.32	22.18	1012.63	(10-20]	[0-10]
708	326160	108	79.45	26.7	398.0	32	GRANDIS	2.64	15.04	1016.46	(10-20]	[0-10]
709	316380	85	71.41	39.4	117.2	21	GRANDIS	3.35	22.25	1012.59	(20-30]	(10-20]
754	156620	4	70.33	28.9	12.0	38	GRANDIS	3.64	21.17	1007.58	(10-20]	[(0-10]

6 CONCLUSÃO

A secagem natural da madeira é, de forma simplificada, um processo de remoção de água da madeira, ou seja, um processo de diminuição do teor de umidade. Este processo ocorre naturalmente, sempre que a madeira é armazenada ao ar livre. Como o armazenamento das toras de madeira em pilhas é considerada a prática comum das organizações que visam a exportação, ou até mesmo a utilização da madeira em processos industriais, pode-se concluir que sempre irá ocorrer variação no teor de umidade enquanto a madeira estiver armazenada – salvaguardando os limites inferiores de perda de umidade.

A variação do teor de umidade da madeira armazenada em pilhas tem, por consequência direta, a variação do peso desta mesma madeira. Por isso, este é um fator é importantíssimo na cadeia de produção de todos os processos industriais que possuem a madeira como matéria-prima. Tal processo é demonstradamente, neste trabalho, um sistema complexo e não-linear, com características variáveis no tempo e, ao mesmo tempo, incertas.

Tais variações na madeira – tanto o teor de umidade quanto o peso – acabam ocasionando perdas irreversíveis na produtividade industrial, e um dos principais motivos destas perdas reside nas rachaduras. As rachaduras são decorrentes, principalmente, da variação do teor de umidade. Além das perdas industriais, outros fatores são decorrentes destas variações do teor de umidade, como a variação do peso da madeira, as perdas logísticas e, não menos importante, as perdas comerciais.

A maioria das organizações brasileiras que armazenam toras de madeira ao ar livre não tem o costume de mensurar o teor de umidade nas pilhas. Em contrapartida, estas mesmas organizações observam de forma tácita esta perda quando verificam que os pesos carregados pelos caminhões que transportam a mesma madeira são drasticamente diminuídos quando comparados aos pesos de quando depositaram esta mesma madeira nos pátios de armazenamento.

Ao propor um modelo de predição baseado em *Machine Learning* utilizando tão somente o peso da madeira – que pode ser mensurado diretamente das balanças rodoviárias, ou até mesmo dos documentos fiscais de transporte das cargas de madeira – além dos dados específicos da madeira, e outros dados meteorológicos do local de armazenamento,

busca-se contribuir diretamente com a solução para este problema de pesquisa de forma extremamente simples e, dessa forma, aumentar o potencial de uso da mesma madeira que é utilizada atualmente.

Por isso, possibilitar a predição da perda do teor de umidade através da variação do peso da madeira, tornará otimizados os processos produtivos que utilizam a madeira, pois as rachaduras nas toras passarão a serem previstas, possibilitando, assim, ações efetivas tanto em tratamentos corretivos quando preventivos. Além disso, ainda que o presente trabalho não possua um *background* ambiental, um modelo que apresente o cálculo da perda de peso das toras de madeira em função do tempo de armazenamento pode fornecer estimativas precisas para a cadeia industrial a ponto de não ser mais necessária a derrubada excessiva de árvores para a obtenção da mesma quantidade final de madeira bruta.

Através deste trabalho, pode-se concluir que o modelo desenvolvido utilizando o algoritmo *Random Forest*, aplicado com os hiperparâmetros ajustados com o *best tuning* sobre o conjunto de dados balanceado, pode ser considerado o modelo mais adequado à solução do presente problema de pesquisa aplicada utilizando métodos de *Machine Learning*.

Pensando em dar prosseguimento a este tema de pesquisa, sugere-se a aplicação de algoritmos baseados em aprendizado profundo (*deep learning*) sobre conjuntos de dados maiores¹; a aferição de quais características – tanto pertencentes à classe meteorológica quanto à classe operacional – podem ser retiradas do conjunto de dados de treinamento sem que haja prejuízo ao desempenho dos modelos; e o *benchmarking* da solução que se utiliza de *Machine Learning* com a melhor solução possível – e disponível no mercado – que não se utiliza de métodos computacionais, para fins de comparação em termos de assertividade.

¹ A principal razão pela qual tal classe de algoritmos não foi utilizada foi a de que o conjunto de dados disponíveis não possui instâncias suficientes para inferir um bom resultado na predição.

REFERÊNCIAS

- [1] Adam, S. P., Alexandropoulos, S.-A. N., Pardalos, P. M., and Vrahatis, M. N. (2019). No free lunch theorem: A review. *Approximation and optimization: Algorithms, complexity and applications*, pages 57–82.
- [2] Allen, J. C. and Barnes, D. F. (1985). The causes of deforestation in developing countries. *Annals of the association of American Geographers*, 75(2):163–184.
- [3] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., and Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and unsupervised learning for data science*, pages 3–21.
- [4] Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.
- [5] Ângelo, H., Almeida, A. N. d., and Serrano, A. L. M. (2009). Determinantes da demanda de madeira em toras para celulose no brasil. *Revista Scientia Forestalis*, 37(84).
- [6] Barbu, M. C. and Tudor, E. M. (2022). State of the art of the chinese forestry, wood industry and its markets. *Wood Material Science & Engineering*, 17(6):1030–1039.
- [7] Bardak, S. and Bardak, T. (2019). Moisture content prediction of wood drying process using fuzzy logic. *Acta Scientiarum Polonorum Silvarum Colendarum Ratio et Industria Lignaria*, 18(1):71–75.
- [8] Batista, G. E. d. A. P. et al. (2003). *Pré-processamento de dados em aprendizado de máquina supervisionado*. PhD thesis, Universidade de São Paulo.
- [9] Batista, G. E. d. A. P. A. and Monard, M. C. (1997). *Um ambiente de avaliação de algoritmos de aprendizado de máquina utilizando exemplos*. PhD thesis, Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação.
- [10] Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10).

- [11] Berry, M. W., Mohamed, A., and Yap, B. W. (2019). *Supervised and unsupervised learning for data science*. Springer.
- [12] Braiek, H. B. and Khomh, F. (2020). On testing machine learning programs. *Journal of Systems and Software*, 164.
- [13] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.
- [14] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- [15] Cortina Orts, A. (2019). Ética de la inteligencia artificial. In *Anales de la Real Academia de Ciencias Morales y Políticas*, pages 379–394. Ministerio de Justicia.
- [16] Daassi-Gnaba, H., Oussar, Y., Merlan, M., Ditchi, T., Géron, E., and Holé, S. (2017). Wood moisture content prediction using feature selection techniques and a kernel method. *Neurocomputing*, 237:79–91.
- [17] de Castro, L. N. and Ferrari, D. G. (2017). *Introdução a mineração de dados*. Saraiva Educação SA.
- [18] Denig, J., EM, W., et al. (1982). Estimating air-drying moisture content losses for red oak and yellow-poplar lumber. *The Forest Products Laboratory*.
- [19] Dorigato, A., Fredi, G., and Pegoretti, A. (2019). Thermo-mechanical behavior of novel wood laminae-thermoplastic starch biodegradable composites with thermal energy storage/release capability. *Frontiers in Materials*, 6:76.
- [20] Eberhardt, J. J. (2015). Bayesian spam detection. *Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal*, 2(1):2.
- [21] Efron, B. (2013). Bayes’ theorem in the 21st century. *Science*, 340(6137):1177–1178.
- [22] El Naqa, I. and Murphy, M. J. (2015). What is machine learning? In *machine learning in radiation oncology*, pages 3–11. Springer.
- [23] Erber, G., Kanzian, C., Stampfer, K., et al. (2012). Predicting moisture content in a pine logwood pile for energy purposes. *Silva Fennica*, 46(4):555–567.
- [24] Estuqui Filho, C. A. (2006). A durabilidade da madeira na arquitetura sob a ação dos fatores naturais: estudo de casos em Brasília. Master’s thesis, Faculdade de Arquitetura e Urbanismo - Universidade de Brasília.

- [25] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.
- [26] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- [27] García, S., Fernández, A., Luengo, J., and Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, 13:959–977.
- [28] Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.
- [29] Gerhards, C. C. (1982). Effect of moisture content and temperature on the mechanical properties of wood: an analysis of immediate effects. *Wood and Fiber Science*, pages 4–36.
- [30] Gerwing, J., Vidal, E., Veríssimo, A., and Uhl, C. (2001). Rendimento no processamento de madeira no estado do Pará. *Série Amazônia*, 18.
- [31] Gigler, J. K., van Loon, W. K., Van den Berg, J., Sonneveld, C., and Meerdink, G. (2000). Natural wind drying of willow stems. *Biomass and Bioenergy*, 19(3):153–163.
- [32] Goldschmidt, R. and Passos, E. (2005). *Data mining*. Gulf Professional Publishing.
- [33] Harrington, P. (2012). *Machine learning in action*. Simon and Schuster.
- [34] Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360.
- [35] Hodges Jr, J. L. and Lehmann, E. L. (2011). Rank methods for combination of independent experiments in analysis of variance. In *Selected works of EL Lehmann*, pages 403–418. Springer.
- [36] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.
- [37] Hongjun Lu, A. (1995). A connectionist approach to data mining. In *Proceedings of the 21 VLDB Conference, Zurich, Switzerland*, pages 81–106.
- [38] Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

- [39] Johnson, J. M. and Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.
- [40] Júnior, A. S. and Alves, J. E. (2019). Fatores intervenientes no armazenamento de laminados. *Gepros: Gestão da Produção, Operações e Sistemas*, 14(5):190.
- [41] Juvenal, T. L. and Mattos, R. L. G. (2002). O setor florestal no brasil e a importância do reflorestamento.
- [42] Kengen, S. (2001). a política florestal brasileira: uma perspectiva histórica [brazilian forest policy: a historical perspective]. *1º Simpósio Ibero-americano de gestão e economia Florestal, Porto Seguro, Brazil. Instituto de Pesquisas e estudos Florestais.(2001). Retrieved May, 4:2007.*
- [43] Kofman, P. D. and Kent, T. (2009). Long term storage and seasoning of conifer energy wood. *Coford connects–Harvesting/Transportations*, 20:1–4.
- [44] Kröll, K. (1959). Selbsttätiges steuern und regeln der trockner. In *Trockner und Trocknungsverfahren*, pages 548–563. Springer.
- [45] Kumar, V. and Minz, S. (2014). Feature selection: a literature review. *SmartCR*, 4(3):211–229.
- [46] LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18):2395–2399.
- [47] Liashchynskyi, P. and Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*.
- [48] Lima, R., da Silva Cardoso, G., de Proença, G., and da Costa, W. G. (2017). Influência do tempo de armazenamento (tpc) da madeira no aceite de cavacos de eucalipto para a fabricação de polpa. In *Congresso Internacional de Celulose e Pape*, page 8. Associação Brasileira Técnica de Celulose e Papel – ABTCP.
- [49] Magalhães., V., Lucca., G., Bicho., A., and Borges., E. (2022). On the methods to predict moisture content on wood: A literature review. In *Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 521–528. INSTICC, SciTePress.
- [50] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386.
- [51] Mazzoni, P., Andersen, R. A., and Jordan, M. I. (1991). A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences*, 88(10):4433–4437.

- [52] Menardi, G. and Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, 28:92–122.
- [53] Mitchell, P. H. (2019). Development of an air drying index for lumber. *BioResources*, 14(3):6657–6665.
- [54] Mitchell, T. (1997). Machine learning.
- [55] Monard, M. C. and Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, 1(1):32.
- [56] Morin, M., Gaudreault, J., Brotherton, E., Paradis, F., Rolland, A., Wery, J., and Laviolette, F. (2020). Machine learning-based models of sawmills for better wood allocation planning. *International Journal of Production Economics*, 222.
- [57] Murphy, K. P. et al. (2006). Naive bayes classifiers. *University of British Columbia*, 18(60):1–8.
- [58] Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, 4:51–62.
- [59] Noriega, L. (2005). Multilayer perceptron tutorial. *School of Computing. Staffordshire University*, 4:5.
- [60] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- [61] Potencial Florestal (2020). A utilização do cavaco de madeira de eucalipto para geração de energia. <https://potencialflorestal.com.br/cavaco-de-eucalipto-para-a-energia>. Acesso: junho de 2023.
- [62] Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R., and Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, 22(8):488–502.
- [63] Quinlan, J. R. (1994a). C4.5: Programs for machine learning. *Machine Learning*, 16(3).
- [64] Quinlan, J. R. (1994b). Induction of decision trees. *Machine Learning*, 1(1):81–106.
- [65] Raitila, J., Heiskanen, V.-P., Routa, J., Kolström, M., and Sikanen, L. (2015). Comparison of moisture prediction models for stacked fuelwood. *BioEnergy Research*, 8(4):1896–1905.

- [66] Rauber, T. W. (2005). Redes neurais artificiais. *Universidade Federal do Espírito Santo*, 29.
- [67] Ray, S. (2019). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMIT-Con)*, pages 35–39. IEEE.
- [68] Ren, F. and Quan, C. (2012). Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing. *Information Technology and Management*, 13:321–332.
- [69] Rezende, R. N., Lima, J. T., de Ramos, L. E., Faria, A. L. R., et al. (2010). Secagem ao ar livre de toras de eucalyptus grandis em lavras, mg. *Cerne*, 16:41–47.
- [70] Rodríguez-Fdez, I., Canosa, A., Mucientes, M., and Bugarín, A. (2015). STAC: a web platform for the comparison of algorithms using statistical tests. In *Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*.
- [71] Russell, S. and Norvig, P. (2002). *Artificial intelligence: a modern approach*. Prentice Hall.
- [72] Samistraro, G. and Hart, P. W. (2012). From tree cutting to pulping: the impact of storage time on pulp mill economics. *TAPPI JOURNAL*, 11(11):43–49.
- [73] Simpson, W. T. (2000). *Estimates of air drying times for several hardwoods and softwoods*, volume 121. US Department of Agriculture, Forest Service, Forest Products Laboratory.
- [74] Simpson, W. T. (2004). *Estimating air drying times of lumber with multiple regression*, volume 293. US Department of Agriculture, Forest Service, Forest Products Laboratory.
- [75] Simpson, W. T. and Wang, X. (2003). Estimating air drying times of small-diameter ponderosa pine and douglas-fir logs. *Res. Pap. FPL-613. Madison, WI: US Dept. of Agriculture, Forest Service, Forest Products Laboratory, 2003. 14 pages.*, 613.
- [76] SNA (2019). Exportação de tora de eucalipto dobra com demanda chinesa. <https://www.sna.agr.br/exportacao-de-tora-de-eucalipto-dobra-com-demanda-chinesa/>. Acesso: outubro de 2021.
- [77] Souza, G. S. A. d. (2019). *Aprendizado de máquina em aplicações de manejo florestal*. PhD thesis, Universidade Federal de Viçosa.

- [78] Speltz, F. R. (2015). Avaliação do potencial de uso de eucalyptus saligna smith e eucalyptus dunnii maiden para produção de dormentes. B.S. thesis, Universidade Tecnológica Federal do Paraná.
- [79] St, L., Wold, S., et al. (1989). Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272.
- [80] Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.
- [81] Suykens, J., De Brabanter, J., Lukas, L., and Vandewalle, J. (2002). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48(1):85–105.
- [82] Tan, J., Yang, J., Wu, S., Chen, G., and Zhao, J. (2021). A critical look at the current train/test split in machine learning. *arXiv preprint arXiv:2106.04525*.
- [83] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- [84] Thörnqvist, T. (1985). Drying and storage of forest residues for energy production. *Biomass*, 7(2):125–134.
- [85] Tomczak, A., Grodziński, G., Jakubowski, M., Jelonek, T., and Grzywiński, W. (2018). Effects of short-term storage method on moisture loss and weight change in beech timber. *Croatian Journal of Forest Engineering: Journal for Theory and Application of Forestry Engineering*, 39(1):35–43.
- [86] Tomczak, K., Tomczak, A., and Jelonek, T. (2020). Effect of natural drying methods on moisture content and mass change of scots pine roundwood. *Forests*, 11(6):668.
- [87] Tsoumis, G. et al. (1991). *Science and technology of wood: structure, properties, utilization*, volume 115. Van Nostrand Reinhold New York.
- [88] Vieira, P. A., Buainain, A. M., and Figueiredo, E. V. C. (2019). O brasil alimentará a china ou a china engolirá o brasil? *Revista Tempo Do Mundo*, 2(1):51–82.
- [89] Visser, R., Berkett, H., and Spinelli, R. (2014). Determining the effect of storage conditions on the natural drying of radiata pine logs for energy use. *New Zealand Journal of Forestry Science*, 44(1):1–8.
- [90] Walker, J. C. (2006). *Primary wood processing: principles and practice*. Springer Science & Business Media.

- [91] Wen, S., Deng, M., and Inoue, A. (2012). Moisture content prediction of wood drying process using svm-based model. *Int. J. Innov. Comput. Infomation Control*, 8:4083–4093.
- [92] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82.
- [93] Wu, H. and Avramidis, S. (2006). Prediction of timber kiln drying rates by neural networks. *Drying Technology*, 24(12):1541–1545.
- [94] Ying, Y., Mursitama, T., et al. (2021). Effectiveness of the news text classification test using the naïve bayes’ classification text mining method. In *Journal of Physics: Conference Series*, volume 1764. IOP Publishing.
- [95] Zhang, X.-D. (2020). Machine learning. In *A Matrix Algebra Approach to Artificial Intelligence*, pages 223–440. Springer.