MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Especialização em Ciência de Dados

UTILIZAÇÃO DE DADOS PELA AUDITORIA PARA PREVENÇÃO DE FRAUDES NO PROCESSO DE RECEBIMENTO DE MADEIRA DE FOMENTO

Gustavo Perpétuo de Andrade

João Monlevade, MG
2024

Gustavo Perpétuo de Andrade

UTILIZAÇÃO DE DADOS PELA AUDITORIA PARA PREVENÇÃO DE FRAUDES NO PROCESSO DE RECEBIMENTO DE MADEIRA DE FOMENTO

Trabalho de conclusão de curso apresentado ao curso de Ciência de Dados do Instituto de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto, como parte dos requisitos necessários para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Harlei Miguel de Arruda Leite

João Monlevade, MG

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

A553u Andrade, Gustavo Perpétuo de.
Utilização de dados pela auditoria para prevenção de fraudes no processo de recebimento de madeira de fomento. [manuscrito] / Gustavo Perpétuo de Andrade. - 2024.
33 f.: il.: color., tab..

Orientador: Prof. Dr. Harlei Miguel de Arruda Leite.
Produção Científica (Especialização). Universidade Federal de Ouro Preto. Departamento de Engenharia de Produção.

1. Análise de dados. 2. Madeira. 3. Papel - Indústria. 4. Redes neurais (Computação). 5. Aprendizado de máquina. 6. Fraude. 7. Auditoria. I. Leite, Harlei Miguel de Arruda. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.85:674.06

Bibliotecário(a) Responsável: Sione Galvão Rodrigues - CRB6 / 2526

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE ENGENHARIA DE PRODUCAO - ICEA

FOLHA DE APROVAÇÃO

Gustavo Perpétuo de Andrade

Utilização de dados pela auditoria para prevenção de fraudes no processo de recebimento de madeira de fomento

Trabalho de conclusão de curso apresentado ao curso de Especialiazação em Ciência de Dados da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Especialista em Ciência de Dados

Aprovada em 19 de junho de 2024

Membros da banca

Dr. Harlei Miguel de Arruda Leite - Orientador - Instituto Tecnológico de Aeronáutica
Dra. Luciana Cerqueira Souza Solimani - Celulose Nipo Brasileira - CENIBRA
Dr. Matheus Nohra Haddad - Universidade Federal de Ouro Preto

Harlei Miguel de Arruda Leite, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 16/07/2024

Documento assinado eletronicamente por **Thiago Augusto de Oliveira Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 17/07/2024, às 16:29, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0741339** e o código CRC **BA5D47E0**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0817 - www.ufop.br

Aos dedicados profissionais da auditoria interna, esta dedicação é um tributo ao incalculável valor que agregam à nossa organização, desbravando os intricados caminhos da transparência e da integridade com destemor e diligência. Que cada linha deste trabalho seja uma homenagem ao compromisso e profissionalismo que demonstram diariamente, garantindo a solidez e a confiabilidade que tanto prezamos. Com profunda gratidão e respeito.

Agradecimentos

Primeiramente, gostaria de expressar minha gratidão a DEUS, cuja presença e bênçãos me fortaleceram e guiaram ao longo de todo este processo.

Gostaria de agradecer à minha esposa Silaine, cujo apoio incansável e compreensão foram fundamentais para minha perseverança e conquista. Ao meu amado filho Bernardo, por sua paciência infinita e por compreender os momentos em que precisei me ausentar para me dedicar aos estudos. Aos meus pais Genésio e Vilza, pilares inabaláveis em minha vida, que desde o início me ensinaram os valores da educação, do trabalho árduo e da determinação.

Agradeço à CENIBRA e à equipe da Auditoria Interna por confiar em meu potencial e por proporcionar o ambiente para o desenvolvimento deste trabalho. A oportunidade de aplicar meus conhecimentos e habilidades em um contexto desafiador e enriquecedor foi uma verdadeira honra.

À Universidade Federal de Ouro Preto (UFOP) e aos professores que ao longo desta jornada foram fontes de inspiração, conhecimento e orientação, expresso minha profunda gratidão.

"Os auditores internos desempenham um papel crucial na promoção da transparência, integridade e responsabilidade dentro das organizações, contribuindo para a confiança dos stakeholders e o sucesso a longo prazo"

Instituto dos Auditores Internos (IIA)

Resumo

O abastecimento de madeira para as indústrias de papel e celulose é uma atividade essencial que requer uma logística eficiente para garantir a continuidade da produção. Nesse contexto, a medição precisa do volume de madeira entregue é crucial para garantir transações justas e evitar perdas financeiras devido a práticas fraudulentas. Este trabalho aborda o desenvolvimento de um modelo de classificação de fraudes no processo de medição de madeira, com foco na distinção entre medições legítimas e fraudulentas. Para atingir esse objetivo, foram utilizadas técnicas de aprendizado de máquina, incluindo regressão linear e redes neurais, para analisar e classificar os dados de medição. Os resultados indicaram que a rede neural MLP superou a regressão linear em termos de precisão e capacidade de detecção de fraudes. A matriz de confusão revelou uma baixa taxa de erros de classificação, validando a eficácia do modelo proposto. Este estudo demonstra o potencial das técnicas de aprendizado de máquina para melhorar a eficiência e a precisão na detecção de fraudes no processo de medição de madeira, contribuindo para aprimorar a gestão e a integridade das operações florestais. Como trabalhos futuros, sugere-se a expansão da análise com a inclusão de novas variáveis, bem como a identificação de outros métodos de aprendizado de máquina para aprimorar o desempenho do modelo de classificação de fraudes.

Palavras-chaves: madeira, papel e celulose, medição, fraude, aprendizado de máquina, regressão linear, redes neurais, detecção de fraudes.

Abstract

The supply of wood for the paper and cellulose industries is an essential activity that requires efficient logistics to ensure continuity of production. In this context, accurate measurement of the volume of wood delivered is crucial to ensure fair transactions and avoid financial losses due to fraudulent practices. This work addresses the development of a fraud classification model in the wood measurement process, focusing on the distinction between legitimate and fraudulent measurements. To achieve this goal, machine learning techniques, including linear regression and neural networks, were used to analyze and classify the measurement data. The results indicated that the MLP neural network outperformed linear regression in terms of accuracy and fraud detection capabilities. The confusion matrix revealed a low rate of classification errors, validating the effectiveness of the proposed model. This study demonstrates the potential of machine learning techniques to improve efficiency and accuracy in detecting fraud in the wood measurement process, contributing to improving the management and integrity of forestry operations. As future work, it is suggested to expand the analysis with the inclusion of new variables, as well as the identification of other machine learning methods to improve the performance of the fraud classification model.

Keywords: wood, pulp and paper, measurement, fraud, machine learning, linear regression, neural networks, fraud detection.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

DAP Diâmetro à altura do peito (1,30m)

Dmin Dâmetro mínimo para comercialização da madeira

Fe Fator de empilhamento

FSC *Forest Stewardship Council*

m³ Metro Cúbico

NTM Nota de Transporte de Madeira

PU Peso Umido

RPV Relação Peso / Volume

ST Metro Estéreo

UFOP Universidade Federal de Ouro Preto

Sumário

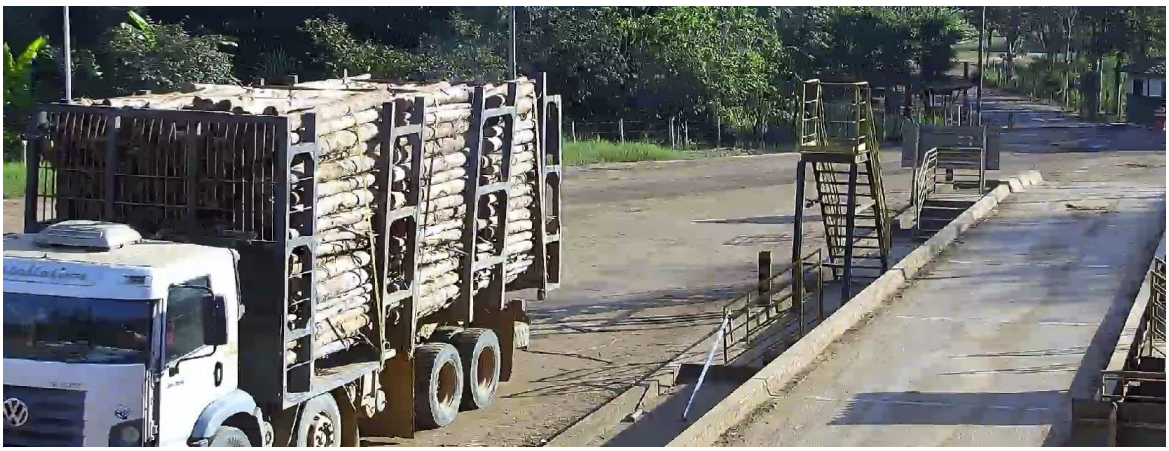
1 Introdução

O abastecimento de madeira até as fábricas de celulose é uma das etapas do processo produtivo e que depende de uma logística eficiente para manter a continuidade da produção. O transporte é realizado principalmente por meio de rodovias e ferrovias, assegurando que a matéria-prima chegue de forma adequada e no tempo previsto.

Para calcular o volume de madeira transportada e entregue, existem diversos tipos de dispositivos de medição, como escâneres a *laser* (*logmeter*), sistemas de fotogrametria, equipamentos de ultrassom, além dos métodos manuais de medição. Cada um possui suas próprias vantagens e limitações, sendo escolhidos de acordo com as necessidades específicas da operação, nível de precisão requerido e custo de aquisição.

Atualmente a CENIBRA dispõe de balança de pesagem, como mostra a Figura 1, na fábrica e nos pátios intermediários para auxiliar na identificação do peso da madeira e cálculo do RPV (Relação Peso / Volume), e adota duas modalidades distintas de medição de acordo com o local de entrega, sendo que na unidade fabril utiliza o equipamento Logmeter com uso de tecnologia de escaneamento a laser e nos pátios intermediários realiza a medição manual da madeira.

Figura 1 – Balança de Pesagem.



Fonte: Autor (2024).

A madeira utilizada no processo pode ser proveniente de duas principais fontes: plantios próprios e plantios de fomento florestal. No modelo de fomento florestal, objeto deste trabalho, produtores rurais são incentivados a plantar, cortar, baldear e entregar a madeira no local especificado pela empresa. Em 2019, aproximadamente 1,6 milhão de pessoas foram beneficiadas por programas de fomento no setor florestal (IBA, 2019).

A adoção de medição manual do volume de madeira é sujeita a erros de subjetividade, pois é um processo em que se utiliza como instrumento réguas para definição de altura, comprimento e largura das pilhas de madeira armazenadas nos caminhões e todas as medições são realizadas por pessoas, como mostra a Figura 2, sendo ainda suscetível a fraudes.

Figura 2 – Caminhão de Fomento - Medição com Régua.

Fonte: Autor (2024).

Este trabalho focaliza na adoção de redes neurais para identificação de variáveis e padrões intrincados nos conjuntos de dados de medição do metro cúbico (m³) de madeira, permitindo classificar e indicar tendências ou indícios de fraude. O processo de descoberta de conhecimento em bancos de dados é uma interação complexa de várias etapas do pré-processamento de dados, seleção de dados, transformação de dados, mineração de dados, avaliação de padrões e conhecimento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

1.1 Objetivo geral

Dada a relevância financeira do metro cúbico de madeira, especialmente em situações em que há riscos de favorecimento de terceiros, é fundamental otimizar a detecção precoce dessas práticas, a fim de mitigar prejuízos. Nos pátios intermediários, a medição manual por régua, conduzida por pessoas, representa uma fonte potencial de vulnerabilidades, suscetíveis a erros inadvertidos ou até mesmo a fraudes deliberadas.

Este trabalho propõe utilizar redes neurais para classificar os registros de medição manual de madeira, oferecendo uma abordagem mais ágil e eficiente. Essa solução pode ser integrada em auditorias contínuas, visando reduzir potenciais prejuízos para a empresa.

1.1.1 Objetivos específicos

Para cumprimento do objetivo geral é essencial atender aos objetivos específicos:

- Identificar e compreender as principais variáveis envolvidas na medição manual de madeira de fomento;
- Determinar a técnica mais eficaz para lidar com essas variáveis, explorando os métodos existentes de redes neurais e desenvolvendo estratégias personalizadas;
- Analisar o desempenho das redes neurais em lidar com diferentes tipos de variáveis envolvidas na medição de madeira de fomento, como características da madeira, instrumentos de medição e pessoas;

1.2 Contribuições

O trabalho propõe o uso de classificação de dados como uma abordagem eficaz para lidar com a complexidade dos processos envolvidos no fomento florestal, onde a atividade de medição ainda é manual. Essa capacidade de análise aprofundada é fundamental para a auditoria interna, permitindo a identificação de indícios de fraudes de forma mais ágil. Essa contribuição é relevante para a literatura sobre auditoria e áreas de mensuração florestal, que dependem de avaliar e calibrar as melhores métricas aplicadas à medição de madeira.

1.3 Organização do Trabalho

Este trabalho é dividido em cinco capítulos que abordam diferentes aspectos do tema em questão. O capítulo atual é a introdução, que mostra uma visão geral do trabalho e sua estrutura. No Capítulo 2 é realizada uma revisão da literatura abrangente, explorando estudos relevantes e oferecendo uma base teórica sólida. No Capítulo 3 é apresentada a metodologia adotada para construção do modelo. No Capítulo 4 os resultados obtidos com os modelos são apresentados e no Capítulo 5 as considerações finais e trabalhos futuros são apresentados.

2 Revisão da Literatura

Esta revisão de literatura tem como objetivo proporcionar uma compreensão abrangente do fomento florestal, detalhando os diversos métodos de medição de madeira e suas respectivas características. Além disso, busca demonstrar as vulnerabilidades inerentes aos mecanismos de medição utilizados, estabelecendo uma correlação entre as fragilidades e indícios de registros fraudulentos.

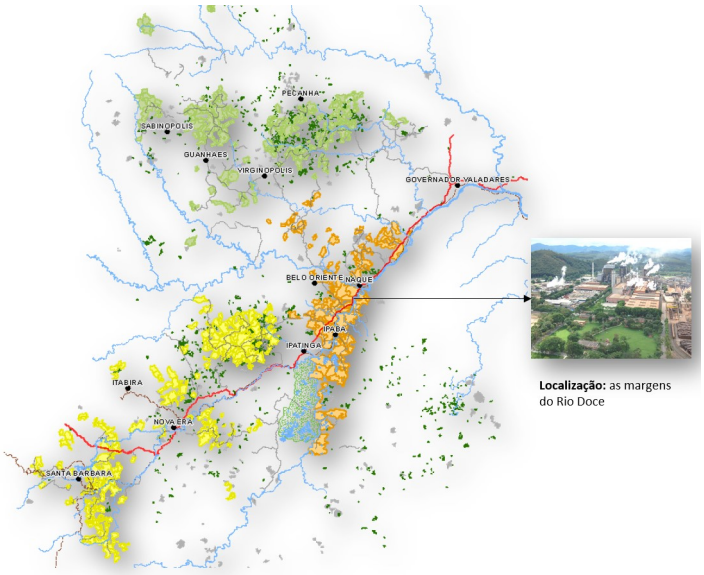
2.1 O fomento florestal de eucalipto

O fomento florestal desempenha um papel estratégico ao conectar os produtores rurais à cadeia produtiva, proporcionando benefícios econômicos, sociais e ambientais. Como resultado, é obtido uma fonte adicional de renda (SIQUEIRA; CANTO; MORAES, 2004).

Por meio de contratos, a CENIBRA facilita o acesso aos insumos, doando mudas e fornecendo assistência técnica. Elas financiam os serviços e insumos para a formação do plantio, permitindo que o produtor implante ou reforme sua floresta sem nenhum desembolso inicial. Os valores financiados são convertidos em madeira e pagos somente na entrega da madeira.

Dessa forma, o custo de implantação florestal para as empresas é reduzido, uma vez que elas não precisam aplicar capital financeiro na aquisição de novas áreas para expansão de suas bases florestais. Atualmente são 21,4 mil hectares (ha) plantados, distribuídos em 78 munícipios, que compreendem três Regionais distintas, conforme mostra a Figura 3.

Figura 3 – Distribuição Fomento Florestal.



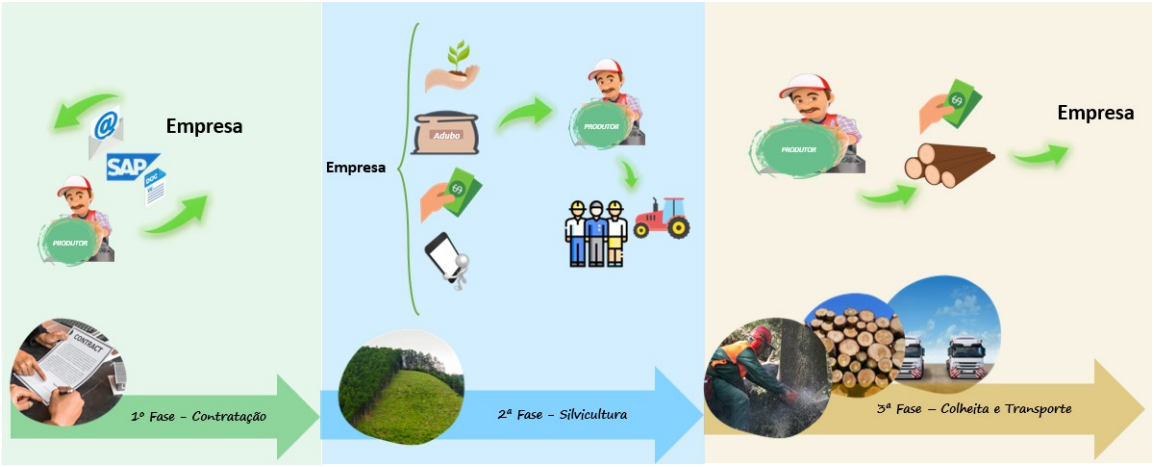
Fonte: Mapa disponibilizado pela Área de Geoprocessamento da CENIBRA (2024).

2.1.1 Fases do processo de fomento

A Figura 4 ilustra de maneira abrangente as diversas fases do processo de fomento florestal, desde a contratação inicial com o produtor até a entrega final da madeira. Cada etapa é meticulosamente delineada, proporcionando uma visão integrada do ciclo das florestas de eucalipto. Nesse contexto, Azevedo e Leite (2024) afirmam que os avanços da tecnologia florestal, aliados à moderna postura socioambiental de grande número de empresas, têm permitido à silvicultura brasileira um lugar de destaque entre as atividades rurais, alinhado às diretrizes do desenvolvimento sustentável.

Este trabalho será realizado na última etapa do fluxo de produção florestal, concentrando-se na medição da madeira de fomento entregue à CENIBRA. Nessa fase crucial, a precisão na medição é fundamental para assegurar a quantidade real de madeira entregue, bem como para cumprir rigorosamente os compromissos contratuais, tanto para o vendedor, quanto o comprador, visando minimizar possíveis fraudes.

Figura 4 – Fases Processo Fomento.

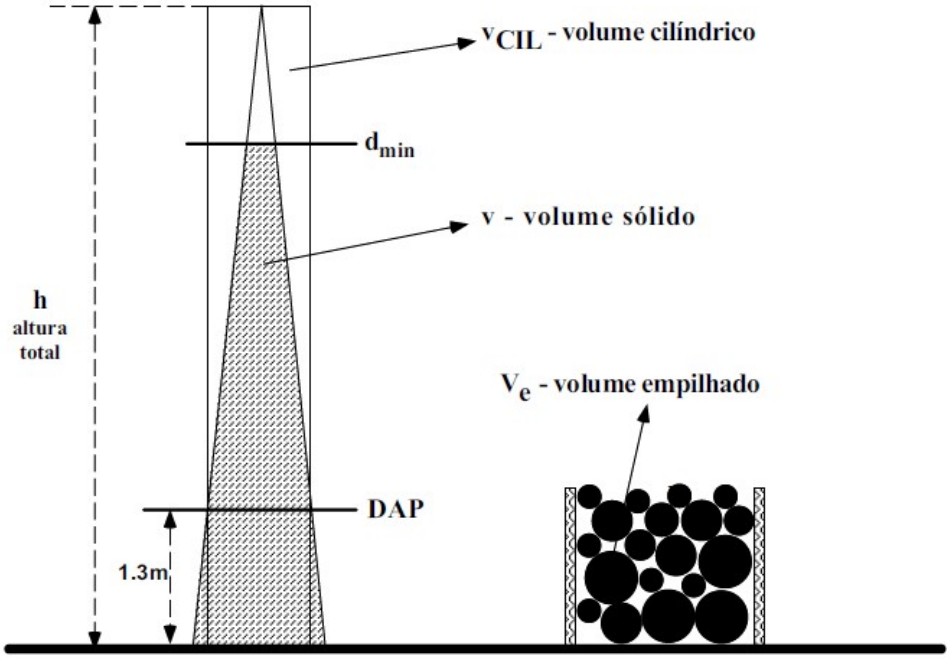


Fonte: Figura disponibilizada pela Área de Fomento da CENIBRA (2024)

2.2 Medição Volumétrica da madeira

Para melhor entendimento do processo de medição da madeira, é preciso compreender todos os fatores considerados para definição do volume de madeira empilhada. A cubagem da madeira pode ser realizada por empilhamento, ou medição de toras individuais, como mostra a Figura 5. No empilhamento, como mostra a Figura 6, as toras são arranjadas em pilhas e o volume é calculado com base nas dimensões da pilha e no fator de empilhamento (MACHADO, 2002).

Figura 5 – Representação Gráfica dos modos de se medir o volume de madeira.



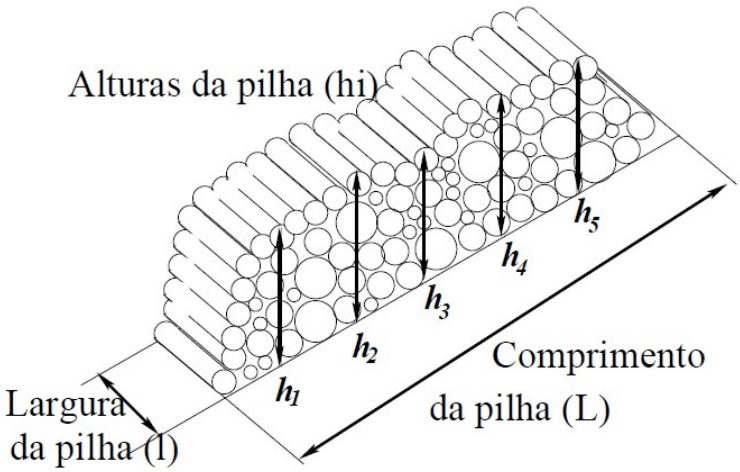
Fonte: (QUANTITATIVOS, 2002).

Para cálculo do volume da pilha são necessárias várias medidas de altura e largura das pilhas e o comprimento das toras, que geralmente é constante e pode ser representado pela fórmula:

$$VolumeEmpilhado = hi \times L \times l \tag{2.1}$$

hi média das alturas da pilha, *L* média das larguras da pilha, e *l* comprimento tora.

Figura 6 – Representação gráfica de uma pilha de madeira e da medição de suas dimensões.



Fonte: (QUANTITATIVOS, 2002).

2.2.1 Volume Estéreo (St)

O volume estéreo de madeira é uma medida que considera o espaço total ocupado por uma pilha de madeira, como mostra a Figura 7, incluindo tanto a madeira sólida quanto os espaços vazios entre os troncos. Essa medida é amplamente utilizada para estimativas rápidas e práticas em contextos em que a madeira é empilhada de maneira irregular, como em caminhões e pátios de fábrica. De acordo com Bertola (2003), o volume estéreo é expresso em (st), onde 1 (um) estéreo corresponde a 1 (um) metro cúbico de espaço ocupado pela combinação de madeira e vazios.

Figura 7 – Pilha Madeira – Medição Estéreo, considera-se a parte sólida e espaços vazios.

Fonte: Autor (2024).

A medição de madeira por volume estéreo pressupõe-se como simples, contudo, não justa, conforme Woodtech (2007), que cita que tal medida considera espaços vazios, buracos na carga, canoas e gaiolas, conforme mostram as Figuras 8 e 9, além de ser intensiva de operações manuais, não auditável e vulnerável a fraudes.

Figura 8 – Detalhe de uma “gaiola” ocasionada por uma tora oblíqua ao alinhamento da pilha.

Fonte: (QUANTITATIVOS, 2002).

Figura 9 – Vista da carroceria de um caminhão com vários “espaços” ocasionados pelo mal empilhamento acidental ou proposital.

Fonte: (QUANTITATIVOS, 2002).

2.2.2 Fator de Empilhamento (Fe)

O fator de empilhamento (Fe) é utilizado para retirada dos espaços vazios contido no volume estéreo quando da conversão para volume sólido (m^3), chegando a um resultado real da madeira disponível para o processo. O fator de empilhamento também é utilizado para diminuir vícios de carregamento e arrumação da carga, conforme destacado.

A utilização de fatores de conversão para quantificar o volume sólido de madeira na pilha é problemática devido às características físicas das toras e à forma como a madeira é empilhada, sendo a variação do fator de empilhamento um dos principais desafios. Em seu estudo, Torquato (1983) destacou que as principais fontes de variação no fator de empilhamento incluem o comprimento, o diâmetro e a forma das toras, o método de empilhamento, a presença de defeitos como tortuosidades e rachaduras nas peças, e o teor de umidade das peças. O autor também observou que a simples modificação do método de empilhamento, de mecânico para manual, pode alterar significativamente o fator de empilhamento.

Além disso, a má disposição das toras pode resultar no "efeito gaiola", que ocorre quando as toras não estão empilhadas paralelamente, aumentando a área vazia da pilha. Este fenômeno é mais provável em pilhas de maior comprimento, devido à maior dificuldade de compactar o empilhamento uniformemente, como mencionado por Keepers (1945).

2.2.3 Volume Sólido (m³)

O volume em metros cúbicos (m³) é uma medida que quantifica exclusivamente a madeira sólida presente em uma pilha, desconsiderando os espaços vazios entre os troncos. Esta medida é fundamental para a comercialização da madeira, pois fornece uma avaliação precisa da quantidade de material disponível para uso. Conforme Bertola (2003), o volume sólido de madeira é essencial para garantir que os cálculos de rendimento e a logística de transporte sejam baseados em dados precisos, refletindo a quantidade real de madeira utilizável.

A conversão do volume estéreo (que inclui espaços vazios) para o volume sólido é feita utilizando um fator de empilhamento. Machado *et al.* (2003) destacam que, embora o fator de empilhamento facilita a conversão do volume estéreo para sólido, ele pode introduzir incertezas nas estimativas devido às variações na forma como a madeira é empilhada e à proporção de espaços vazios na pilha. Portanto, para obter medições precisas do volume em m³, é crucial considerar as condições específicas do empilhamento e aplicar fatores de correção adequados. A CENIBRA adota um fator de empilhamento padrão para cada regional, mediante a avaliação periódica das cargas, sendo revisitado o processo semestralmente.

2.3 Métodos de Medição

A medição do volume de madeira envolve um conjunto complexo de procedimentos e apresenta diversos desafios, uma vez que as toras não são homogêneas e uniformes, a forma de empilhamento e disposição da madeira não permite um alinhamento perfeito e cada madeira apresenta uma densidade diferente, podendo variar a densidade até mesmo de um caminhão para outro, mesmo a madeira sendo do mesmo local.

Atualmente a CENIBRA adota dois métodos, equipamento eletrônico *Logmeter* (automatizado) e medição com Régua (manual), a definição de qual utilizar considera os aspectos de quantidade de madeira processada em cada local e o custo-benefício.

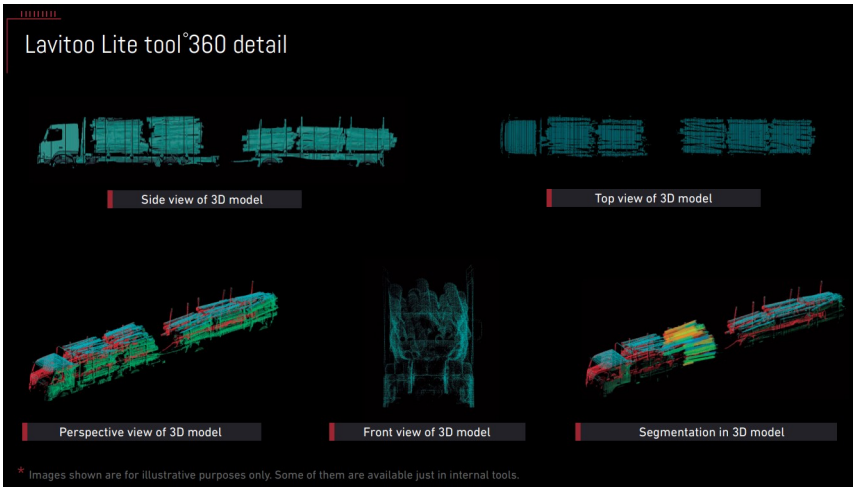
2.3.1 Equipamento *Logmeter* - Automatizado

O *Logmeter*, utilizado na unidade fabril, é um sistema automatizado projetado para realizar medições precisas e estimativas biométricas das toras de madeira enquanto estão carregadas em caminhões de transporte (WOODTECH MEASUREMENT, 2014). Este sistema emprega tecnologia *laser* avançada para alcançar resultados de alta precisão sem a necessidade de intervenção direta de um operador, como mostra as Figuras 10 e 11.

Figura 10 – *Logmeter*.

Fonte: (WOODTECH MEASUREMENT, 2014).

Figura 11 – Logmeter - Visão 360º.



Fonte: (WOODTECH MEASUREMENT, 2014).

O sistema representa uma inovação significativa no campo da medição florestal, pois oferece uma alternativa eficaz e eficiente para a medição manual tradicional (WOODTECH MEASUREMENT, 2014). Ao eliminar a necessidade de intervenção direta do operador, o sistema reduz a possibilidade de erros humanos e aumenta a confiabilidade das medições obtidas.

O *Logmeter* é capaz de medir o volume estéreo ocupado por um feixe de troncos, proporcionando uma avaliação precisa do espaço total ocupado pela madeira, incluindo os vazios entre os troncos. Além disso, determina o comprimento médio dos troncos detectados, garantindo um cálculo exato das dimensões longitudinais médias. Ele também mede o diâmetro médio dos troncos detectados, fornecendo uma visão geral do tamanho transversal médio da madeira. O desvio padrão dos diâmetros dos troncos detectados é calculado, permitindo uma análise da variabilidade dos tamanhos dos troncos dentro do feixe. A ferramenta também delineia o contorno dos troncos na periferia, facilitando a identificação e a medição precisa das bordas do feixe. Finalmente, o *Logmeter* avalia a relação sólido/estéreo na periferia, oferecendo uma métrica importante para entender a densidade e a compactação da madeira nas extremidades do feixe.

Para a madeira de fomento, fornecida por terceiros, a taxa de erro das medições do logmeter na CENIBRA está entre 3 e 5%, podendo alcançar até 7% conforme informado pela área de mensuração florestal da CENIBRA e previsto no contrato de fornecimento do equipamento.

2.3.2 Medição com Régua - Manual

O método manual, com a utilização de régua para medição do volume (m³) de madeira carregada nos caminhões é, sem dúvida, trabalhoso e demorado. Este processo está sujeito a uma série de variáveis que podem introduzir erros significativos, especialmente devido à subjetividade nas medições realizadas por diferentes pessoas. Além disso, a complexidade e a necessidade de intervenção humana tornam este método vulnerável a fraudes, onde medições imprecisas ou intencionais podem resultar em registros incorretos.

Um dos principais desafios desse método é a necessidade de medir o volume empilhado em estéreos e convertê-lo para volume sólido em metros cúbicos. Este processo é fundamental para calcular o volume efetivo de madeira, excluindo os espaços vazios presentes nas pilhas. Mesmo que o fator de empilhamento possua um padrão e definido internamente pela CENIBRA com base em estudos, os gabaritos dos caminhões são passivéis de manipulação e para a realização das medições com régua, sendo 06 medidas de altura e 06 de comprimento, há intervenção humana em todo o processo, conforme mostra a Figura 12.

Segundo Hägglund (2006), a medição manual da madeira empilhada está sujeita a uma alta variabilidade devido à subjetividade e habilidade do operador humano, podendo resultar em resultados inconsistentes e imprecisos.

Figura 12 – Medição Manual pelo recebedor de madeira.

Fonte: Autor (2024).

A Tabela 1 demonstra as medições realizadas pelo recebedor e a Tabela 2 representa o cálculo realizado pelo sistema para definição do volume, levando em consideração as médias das alturas e comprimento, número de lastros do gabarito do caminhão e fator de empilhamento. Qualquer alteração das medidas resultaria em aumento ou diminuição do volume m³.

Tabela 1 – Medições da madeira empilhada.

Item	**Medições Manuais (Pessoas)**						
Altura Pilha	A1	A2	A3	A4	A5	A6	Média
	2,37	2,42	2,40	2,40	2,43	2,45	2,41
Comprimento Lastros	C1	C2	C3	C4	C5	C6	Média
	2,25	2,25	2,25	2,18	2,25	2,25	2,24

Fonte: Autor (2024).

Tabela 2 – Cálculo Volume m³.

Cálculo Sistema ERP - Volume m³				
Altura	**Comprimento**	**Nº Lastros**	**Fe Mec**	**Volume m³**
2,41	2,24	3	1,55	**27,126**

Fonte: Autor (2024).

De forma ilustrativa, como mostra a Figura 13, a comparação visual do mesmo veículo, no mesmo dia e local, revelou que a pilha de maior altura apresentava um volume menor.

Figura 13 – Comparativo visual de empilhamento no mesmo veículo.

Horário: 10:07:47 / Recebedor: YPTO
Volume medido: **21,96m³**
Peso auferido: 19.720kg

Horário: 16:54:10 / Recebedor: XPTO
Volume medido: **28,44 m³**
Peso auferido: 18.810kg

Fonte: Autor (2024).

2.3.3 Método Regressão Linear

A regressão linear é um método estatístico amplamente utilizado para modelar e analisar a relação entre uma variável dependente e uma ou mais variáveis independentes. Sua popularidade deve-se à simplicidade, interpretabilidade e eficiência computacional, sendo uma das primeiras técnicas a serem ensinadas em cursos de estatística e ciência de dados. A origem da regressão linear remonta aos trabalhos de Francis Galton no século XIX, que utilizou a técnica para estudar a hereditariedade. Posteriormente, Karl Pearson e outros matemáticos desenvolveram o conceito de correlação e análise de regressão linear, formalizando os métodos estatísticos que são a base da técnica moderna (STIGLER, 1986).

A regressão linear simples pode ser expressa pela equação

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.2}$$

onde y é a variável dependente, x é a variável independente, β_0 é o intercepto, β_1 é o coeficiente de inclinação e ε é o termo de erro. A extensão para múltiplas variáveis independentes resulta na regressão linear múltipla:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{2.3}$$

(MONTGOMERY; PECK; VINING, 2012).

A regressão linear é utilizada em diversas áreas, incluindo economia, engenharia, biologia e ciências sociais. Exemplos práticos incluem a previsão de preços de imóveis, análise de risco de crédito e estudos epidemiológicos (TIBSHIRANI, 1996). Sua capacidade de fornecer uma relação direta e interpretável entre variáveis a torna uma ferramenta valiosa em análises preditivas e inferenciais.

Apesar de suas vantagens, a regressão linear apresenta limitações. Assumir linearidade entre variáveis pode ser inadequado para dados complexos e não lineares. Além disso, a técnica é sensível a outliers e à multicolinearidade, o que pode afetar a precisão dos coeficientes estimados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Com o advento do aprendizado de máquina, a regressão linear foi integrada em algoritmos mais complexos, como regressão por partes e regularização (Lasso e Ridge). Esses avanços permitem lidar com grandes volumes de dados e melhorar a robustez e a precisão das previsões (TIBSHIRANI, 1996).

A regressão linear continua sendo uma técnica fundamental em ciência de dados, devido à sua simplicidade e eficácia. No entanto, é importante reconhecer suas limitações e complementá-la com outras técnicas quando necessário para obter análises mais precisas e robustas.

2.3.4 Método Redes Neurais MLP

As redes neurais perceptron multicamada (MLP, do inglês *Multilayer Perceptron*) são uma das arquiteturas mais conhecidas e estudadas no campo das redes neurais artificiais. Elas desempenham um papel crucial no aprendizado profundo e são amplamente utilizadas em tarefas de classificação e regressão (GOODFELLOW; BENGIO; COURVILLE, 2016). O conceito de *perceptron* foi introduzido por Frank Rosenblatt em 1958, mas as limitações das redes de camada única foram rapidamente identificadas. A introdução do algoritmo de retropropagação por Rumelhart, Hinton e Williams em 1986 permitiu o treinamento eficaz de redes multicamadas, resolvendo problemas que o perceptron simples não conseguia (RUMELHART; HINTON; WILLIAMS, 1986).

Uma rede neural MLP consiste em pelo menos três camadas de neurônios: a camada de entrada, uma ou mais camadas ocultas e a camada de saída. Cada neurônio em uma camada está conectado a todos os neurônios da próxima camada, e o sinal de entrada é propagado através dessas conexões. A função de ativação, geralmente não linear, é aplicada à saída ponderada de cada neurônio:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.4}$$

onde w_i são os pesos das conexões, x_i são as entradas, b é o bias e f é a função de ativação, como ReLU, sigmoid ou tanh (NIELSEN, 2015).

As redes MLP são utilizadas em diversas áreas, incluindo reconhecimento de padrões, processamento de imagem, previsão de séries temporais e processamento de linguagem natural. Exemplos incluem a classificação de dígitos escritos à mão, previsão de vendas e análise de sentimentos em textos (GOODFELLOW; BENGIO; COURVILLE, 2016). Uma das principais vantagens das MLPs é sua capacidade de modelar relações não lineares complexas entre variáveis. No entanto, elas requerem grande quantidade de dados para treinamento e são computacionalmente intensivas. Além disso, o ajuste dos hiperparâmetros e a escolha da arquitetura adequada podem ser desafiadores.

Recentemente, técnicas como regularização, inicialização de pesos e otimização avançada (como *Adam e RMSprop*) melhoraram significativamente o desempenho das redes MLP. Além disso, a integração com arquiteturas mais complexas, como redes convolucionais e recorrentes, ampliou suas capacidades (KINGMA; BA, 2014); (HEATON, 2018).

As redes neurais MLP são uma ferramenta poderosa em ciência de dados, capazes de resolver problemas complexos que métodos tradicionais não conseguem. Embora possuam desafios inerentes, os avanços contínuos na área de aprendizado profundo continuam a aprimorar seu desempenho e aplicabilidade.

2.3.5 Referências na Literatura

A busca por estudos na literatura acerca de fraudes na medição manual de madeira revelou-se desafiadora devido à escassez de trabalhos específicos sobre esse tema. Embora tenham sido encontrados diversos artigos que abordam fraudes relacionadas ao fator de conversão na medição de madeira, a pesquisa detalhada por fraudes exclusivamente no método manual revelou-se inconclusiva.

A literatura consultada evidencia uma preocupação recorrente com fraudes e erros na medição de volumes de madeira, principalmente relacionados à aplicação incorreta de fatores de conversão que ajustam as medidas de volume. Estes fatores são cruciais para converter medidas de diâmetro e comprimento de árvores em estimativas de volume de toras, sendo frequentemente alvo de manipulações fraudulentas que impactam a precisão e a confiabilidade dos dados.

Apesar da relevância desses estudos, a ausência de trabalhos específicos que investiguem exclusivamente fraudes no método manual de medição de madeira sugere uma lacuna significativa na literatura científica. Isso pode ser atribuído à predominância de métodos automatizados e digitais que têm substituído gradualmente a medição manual, reduzindo assim o interesse e o foco em estudos dedicados a fraudes específicas nesse contexto.

Portanto, para preencher essa lacuna de conhecimento, é crucial realizar pesquisas adicionais que investiguem de maneira específica as práticas fraudulentas no método manual de medição de madeira. Esses estudos não apenas contribuirão para a compreensão mais profunda das vulnerabilidades existentes nesse método, mas também para o desenvolvimento de estratégias e tecnologias que possam mitigar esses riscos e aumentar a confiabilidade das medições de recursos florestais.

3 Metodologia

A metodologia adotada neste trabalho teve como objetivo principal realizar um estudo sobre a classificação de medições manuais de madeira de eucalipto em duas categorias distintas: fraude e não fraude. Com base em análises empíricas e teóricas, procurou-se identificar padrões que permitissem discernir entre medições legítimas e práticas fraudulentas. O objetivo final foi desenvolver estratégias eficazes para classificar as medições de acordo com essa distinção, visando garantir a precisão e a integridade das operações de medição.

3.1 Base de Dados

Um passo essencial e desafiador na análise de dados, é a seleção e extração dos dados. Como Han, Kamber e Pei (2011) afirmam, “a qualidade dos dados extraídos afeta diretamente a qualidade da análise de dados”.

Neste estudo, a coleta dos dados provenientes da base de recebimento de madeira foi obtida por meio do Sistema SAP S/4 Hana e da ferramenta Arbutus *Analyzer*. Os dados selecionados foram então extraídos e armazenados em um arquivo no formato ".xlsx". Posteriormente, essa base foi importada para a plataforma Google Colab, onde foi acessada e processada utilizando a linguagem de programação *Python*. Durante esse processo, os dados foram carregados em um *dataframe*, com a variável de interesse "Fraude"sendo reposicionada como a última coluna.

Para a análise propriamente dita, foram extraídos 572 registros, os quais abrangiam 23 variáveis, referentes ao período de janeiro de 2018 a dezembro de 2019. Esses registros compreendem uma variedade de informações relacionadas às entregas de madeira, como produtividade, densidade, volume estéreo e volume em metros cúbicos, entre outras. A Tabela 3 apresenta as variáveis que compõem a amostra.

Tabela 3 – Descrição das Variáveis de Medição de Madeira.

Nome da Variável	**Descrição**
Usuario	Usuário responsável pela medição
Placa	Placa do veículo que realizou a entrega da madeira
Produtividade	Produtividade do contrato de fomento
Idade	Idade da madeira, considerando data de plantio e corte
Dias_corte	Número e dias entre o corte e a entrega
Alt_01	Medição da primeira altura da pilha
Alt_02	Medição da segunda altura da pilha
Alt_03	Medição da terceira altura da pilha
Alt_04	Medição da quarta altura da pilha
Alt_05	Medição da quinta altura da pilha
Alt_06	Medição da sexta altura da pilha
Comp_01	Medição do primeiro comprimento da pilha
Comp_02	Medição do segundo comprimento da pilha
Comp_03	Medição do terceiro comprimento da pilha
Comp_04	Medição do quarto comprimento da pilha
Comp_05	Medição do quinto comprimento da pilha
Comp_06	Medição do sexto comprimento da pilha
Vol_ST	Volume estéreo
Densidade	Densidade da madeira
Qtde_MU	Peso registrado na balança menos a tara do veículo
RPV	Relação Peso/Volume
Vol_m3	Volume em metro cúbico (m^3)
Fraude	Variável de saída (Fraude ou Não Fraude)

Fonte: Autor (2024).

3.2 Pré-processamento de Dados

Com base no banco de dados coletado no sistema SAP da CENIBRA, foram analisadas as variáveis para a construção do modelo de classificação de fraudes no recebimento de madeira, sendo realizada uma análise descritiva, como mostra a Tabela 4.

Foram realizados diversos pré-processamentos essenciais para preparar os dados antes de aplicá-los aos modelos. Inicialmente, os dados foram carregados a partir de um arquivo Excel utilizando a função "*pd.read_excel()*", que permite importar os dados para um *DataFrame* da biblioteca do *Pandas*. Em seguida, os dados foram separados em variáveis independentes e variável dependente, diferenciando os atributos que serão utilizados para a previsão e a variável que se deseja classificar.

Posteriormente, os dados foram divididos em conjuntos de treinamento e teste utilizando a função "*train_test_split()*"da biblioteca *sklearn*. A divisão foi feita de modo que **80%** dos dados fossem destinados ao treinamento dos modelos e **20%** para teste. Foi utilizado o parâmetro "*random_state=42*"para garantir que a divisão dos dados seja reprodutível, permitindo que os resultados possam ser comparados e replicados em futuras execuções.

Tabela 4 – Análise Descritiva.

Variável	count	mean	std	min	25%	50%	75%	max
Usuario	572.0	5	1	1	5	7	7	8
Placa	572.0	8.4	4.1	1.0	5.0	9.0	13.0	15.0
Produtividade	572.0	417	83	280	347	389	420	562
Idade	572.0	8.7	0.7	7.3	8.2	8.8	9.4	11.0
Dias_corte	572.0	71	21	31	55	68	84	195
Alt_01	572.0	2.45	0.21	1.70	2.38	2.50	2.60	2.76
Alt_02	572.0	2.43	0.21	1.75	2.30	2.47	2.60	2.76
Alt_03	572.0	2.40	0.23	1.70	2.25	2.45	2.60	2.76
Alt_04	572.0	2.42	0.21	1.72	2.30	2.46	2.59	2.76
Alt_05	572.0	2.43	0.21	1.75	2.29	2.47	2.60	2.76
Alt_06	572.0	2.43	0.23	1.63	2.28	2.49	2.60	2.76
Comp_01	572.0	2.24	0.02	2.10	2.24	2.25	2.25	2.25
Comp_02	572.0	2.23	0.02	2.10	2.24	2.25	2.25	2.25
Comp_03	572.0	2.23	0.02	2.14	2.24	2.25	2.25	2.55
Comp_04	572.0	2.23	0.02	2.05	2.24	2.25	2.25	2.25
Comp_05	572.0	2.23	0.02	2.08	2.23	2.25	2.25	2.46
Comp_06	572.0	2.23	0.02	2.11	2.23	2.25	2.25	2.55
Vol_ST	572.0	38.27	3.68	28.32	36.10	38.73	41.34	43.67
Densidade	572.0	507.42	18.74	457.31	492.01	517.55	520.68	529.33
Qtde_MU	572.0	20335	2106	14040	19107	19980	21510	43400
RPV	572.0	779	92	514	712	766	851	1411
Vol_m3	572.0	26.29	2.76	20.09	24.21	26.22	28.34	30.97
Fraude	572.0	0.5	0.5	0.0	0.0	0.5	1.0	1.00

Fonte: Autor (2024).

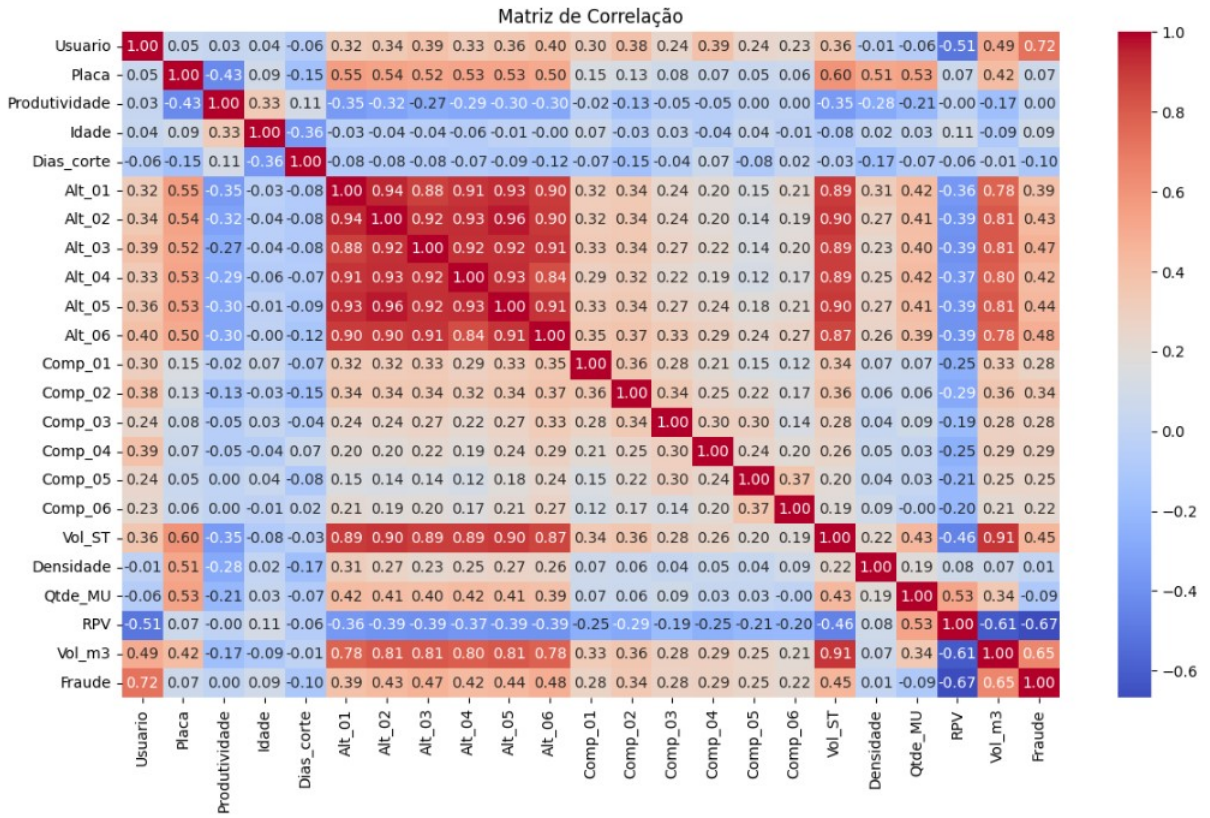
Além disso, foi realizada a normalização dos dados de *features* utilizando o *StandardScaler* da biblioteca *sklearn*. Este processo é fundamental para que todas as variáveis tenham uma média de 0 e uma variância de 1, uniformizando a escala das variáveis e melhorando o desempenho dos algoritmos. A normalização foi aplicada inicialmente ao conjunto de treinamento através do método "*fit_transform()*"e, em seguida, ao conjunto de teste utilizando o método "*transform()*".

Um ponto importante no pré-processamento foi a identificação da variável usuário, que foi transformada para um formato numérico. Essa transformação permitiu a inclusão da variável "usuário"na análise da matriz de correlação. A análise demonstrou que a concentração de registros em um mesmo usuário apresentou uma correlação significativa com a variável de saída fraude. Esse achado indica que a concentração de registros pode ser um indicador relevante de fraude, sugerindo que um volume elevado de transações associadas a um mesmo usuário está correlacionado com um aumento nas chances de ocorrência de fraude. Portanto, essa variável foi considerada uma das mais relevantes para o modelo de classificação.

Esses pré-processamentos são etapas fundamentais na preparação dos dados, assegurando que os modelos de redes neurais possam aprender de maneira eficiente e que os resultados obtidos sejam confiáveis e válidos para avaliação e comparação. Segundo Geron (2019), a preparação adequada dos dados é crucial para o sucesso de qualquer projeto de *machine learning*, pois modelos treinados com dados bem preparados tendem a apresentar desempenho superior e maior generalização em dados não vistos.

A matriz de correlação foi explorada, como mostra a Figura 14, para entender como as variáveis se relacionam entre si, fornecendo *insights* valiosos para identificar as melhores variáveis a serem utilizadas como entrada dos modelos de aprendizado de máquina.

Figura 14 – Matriz de Correlação.



Fonte: Autor (2024).

3.3 Regressão Linear

A regressão linear é uma técnica estatística e de aprendizado de máquina utilizada para modelar a relação entre uma variável dependente, "Fraude"e uma ou mais variáveis independentes, ajustando uma linha reta que minimiza a soma dos erros quadrados entre as previsões do modelo e os valores reais.

Após as etapas de pré-processamento, a implementação foi realizada utilizando a biblioteca scikit-learn, especificamente a classe LinearRegression. O modelo, implementado usando a biblioteca sklearn, como mostra a Figura 15, foi treinado com os dados de treinamento e, em seguida, o desempenho do modelo foi avaliado utilizando o conjunto de validação. A métrica utilizada para avaliação foi a acurácia, que mede a proporção de previsões corretas em relação ao total de previsões.

Figura 15 – Modelo de Regressão Linear.

```
linear_model = LinearRegression()
linear_model.fit(X_train_scaled, y_train)
linear_pred = linear_model.predict(X_val_scaled)
linear_accuracy = accuracy_score(y_val, linear_pred.round())
print('Acurácia do modelo de regressão linear no conjunto de validação:', linear_accuracy)
```

Fonte: Autor (2024).

Ao considerar a natureza da regressão linear, há métricas, assim como a acurácia, que ajudam a entender o quão bem o modelo está ajustado aos dados, como:

- **Precisão**: indica a proporção de instâncias positivas (casos preditos como positivos corretamente) sobre todas as instâncias preditas como positivas pelo modelo. Para a regressão linear, isso pode ser relacionado à precisão das previsões contínuas em relação aos valores reais.
- **Recall**: mede a proporção de instâncias positivas corretamente identificadas pelo modelo entre todas as instâncias que realmente são positivas. Em regressão linear, pode ser interpretado como a capacidade do modelo de capturar corretamente as variações nos dados observados.
- **F1 Score**: é a média harmônica entre precisão e recall e fornece uma medida única da precisão geral do modelo. Em regressão linear, é uma métrica útil para avaliar o equilíbrio entre a precisão das previsões e a cobertura das variações nos dados.
- **Índice Kappa**: avalia a concordância entre as classificações reais e as predições do modelo, considerando a possibilidade de acordos ao acaso. Em regressão linear, pode ser usado para entender a confiabilidade das previsões contínuas do modelo em relação aos valores observados.

3.4 Rede Neural MLP

As redes neurais têm se destacado como uma poderosa ferramenta para diversas tarefas de aprendizado de máquina, como classificação, regressão e reconhecimento de padrões. Entre as várias arquiteturas de redes neurais, o *Perceptron* Multicamadas (MLP, *Multi-Layer Perceptron*) é uma das mais populares e amplamente utilizadas. O MLP é um tipo de rede neural *feedforward* composta por uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada uma dessas camadas desempenha um papel crucial no processamento e na transformação dos dados para alcançar o objetivo desejado.

Neste trabalho as etapas do funcionamento do *Multi-Layer Perceptron* (MLP) foram implementadas para treinar e avaliar a rede neural, como mostra a Figura 16, sendo utilizado:

- Inicialização dos pesos: realizada automaticamente pelo *framework Keras*, que utiliza métodos padrão para inicializar os pesos de maneira aleatória;
- Propagação para Frente (*Forward Propagation*): durante o treinamento e a avaliação do modelo MLP, foi utilizada para calcular as saídas da rede a partir dos dados de entrada, aplicando as funções de ativação em cada neurônio para gerar as previsões finais;
- Cálculo do Erro: calculado utilizando a função de perda *binary_crossentropy*, apropriada para problemas de classificação binária. A perda foi minimizada durante o treinamento, comparando as previsões da rede com os valores reais dos dados de validação;
- Retropropagação (*Backpropagation*): utilizada para calcular os gradientes do erro em relação aos pesos da rede. O *Keras* automaticamente gerencia a retropropagação e a atualização dos pesos durante o treinamento;
- Atualização dos Pesos: atualização iterativamente utilizando o otimizador *Adam*, que é uma escolha popular por sua eficiência e capacidade de adaptabilidade. *Adam* ajusta a taxa de aprendizado durante o treinamento com base nos gradientes calculados, ajudando a rede a convergir mais rapidamente e de forma estável.

3.4.1 Otimização de Hiperparâmetros

O processo de otimização de hiperparâmetros é crucial no treinamento de modelos de aprendizado de máquina, especialmente em redes neurais, cujos parâmetros são definidos antes do início do processo, ao contrário dos parâmetros do modelo que são aprendidos durante.

Figura 16 – Modelo da rede MLP.

```
def build_model(hp):
    model = keras.Sequential()
    for i in range(hp.Int('num_layers', 2, 20)):
        model.add(layers.Dense(units=hp.Int('units_' + str(i),
                                            min_value=32,
                                            max_value=512,
                                            step=32),
                               activation='relu'))
    model.add(layers.Dense(1, activation='sigmoid'))
    model.compile(optimizer=keras.optimizers.Adam(hp.Choice('learning_rate', [1e-2, 1e-3,
                  loss='binary_crossentropy',
                  metrics=['accuracy'])
    return model
```

Fonte: Autor (2024).

O *Keras Tuner* foi utilizado para implementar o *Random Search* e o processo de otimização foi configurado para buscar os melhores hiperparâmetros, como mostra a Figura 17, como o número de camadas ocultas, o número de neurônios em cada camada e a taxa de aprendizado. O *Random Search* foi configurado com um objetivo de maximizar a acurácia de validação (*val_accuracy*), um número máximo de tentativas (*max_trials*) e o número de execuções por tentativa (*executions_per_trial*).

Figura 17 – Otimizador de Hiperparâmetros.

```
# Configurar o otimizador de hiperparâmetros
tuner = RandomSearch(
    build_model,
    objective='val_accuracy',
    max_trials=15,
    executions_per_trial=5,
    directory='keras_tuner',
    project_name='modelo_mlp')

# Pesquisar os melhores hiperparâmetros
tuner.search(X_train_scaled, y_train,
             epochs=5,
             validation_data=(X_val_scaled, y_val))

# Resumo dos melhores hiperparâmetros
tuner.results_summary()

# Obter o modelo com os melhores hiperparâmetros
best_model = tuner.get_best_models(num_models=1)[0]
```

Fonte: Autor (2024).

3.5 Outros Métodos

A escolha do método depende da natureza dos dados, do problema específico (classificação ou regressão) e das necessidades de interpretabilidade e desempenho do modelo, onde a correlação entre métodos, pode ajudar na apuração dos resultados encontrados, tando a título de desempenho, quanto interpretabilidade.

Conhecer outros métodos e comparar os resultados é de suma importância, sendo apresentado abaixo alguns destes de forma sucinta e que o comparativo nos resultados trouxe avaliações interessantes para utilização futura:

- **Regressão Logística**: modelo estatístico usado para prever a probabilidade de uma variável dependente categórica (binária) com base em uma ou mais variáveis independentes, sendo semelhante à regressão linear, mas é aplicada para prever probabilidades de pertencimento a uma classe específica.
- **Árvore de Decisão**: método de aprendizado de máquina supervisionado que divide o espaço de características em regiões retangulares e prevê o valor-alvo em cada região. Pode ser usado tanto para problemas de regressão quanto de classificação.
- **Floresta Aleatória**: técnica de aprendizado de máquina que combina várias árvores de decisão durante o treinamento e faz uma média de suas previsões para melhorar a precisão e evitar overfitting. Muito eficaz para problemas de classificação e regressão, especialmente quando há muitas variáveis preditoras.

4 Resultados

Os resultados encontrados oferecem uma visão abrangente do desempenho dos modelos de regressão linear e Redes Neurais MLP, sendo este último, mais objetivo na classificação de registros de medição com indícios de fraude e não fraude.

4.1 Regressão Linear

Para a regressão linear, é importante notar que a matriz de confusão não é uma métrica de avaliação padrão, pois a regressão linear é usada para prever valores contínuos, não para classificar instâncias em categorias discretas. No entanto, podemos interpretar seu desempenho de maneira mais ampla.

Ao considerar a natureza da regressão linear, focamos principalmente na acurácia do modelo, que nos fornece uma medida geral da capacidade do modelo em explicar a variabilidade nos dados. No caso deste modelo, a acurácia no conjunto de validação foi de aproximadamente 93.04%, indicando que a relação linear entre as características e o rótulo alvo pode explicar cerca de 93.04% da variabilidade nos dados de validação. O percentual de erro associado a este modelo foi de aproximadamente 6.96%.

A Tabela 5 apresenta os resultados específicos para o modelo de regressão linear, destacando a importância dessas métricas na análise do desempenho do modelo:

Tabela 5 – Métricas de desempenho para o modelo de Regressão Linear.

Métrica	**Valor**
Acurácia	0.930
Precisão	0.931
Recall	0.930
F1 Score	0.930
Índice Kappa	0.858

Fonte: Autor (2024).

Embora a regressão linear não tenha fornecido diretamente uma matriz de confusão, a partir dos resultados é possível modelar a relação entre variáveis independentes e dependentes por meio de uma linha reta. Dessa forma, o desempenho da regressão linear foi avaliado com base na qualidade do ajuste do modelo aos dados observados, revelando os seguintes resultados:

- Coeficiente de Determinação (R^2): o valor de R^2 foi de aproximadamente 0.658, o que significa que cerca de 65.8% da variabilidade entre a variável dependente e as independentes nos dados de validação é explicada pelo modelo de regressão linear;

- Erro Quadrático Médio (MSE): o valor de MSE foi de aproximadamente 0.084, indicando uma precisão moderada na previsão dos valores alvo;
- Erro Absoluto Médio (MAE): o valor de MAE foi de aproximadamente 0.198, o que significa que, em média, as previsões do modelo estão a cerca de 0.198 unidades de distância dos valores reais.

Os resultados sugerem que o modelo de regressão linear é capaz de explicar uma parte significativa da variabilidade nos dados, com uma precisão moderada na previsão dos valores alvo, mas que, melhores resultados podem ser obtidos através de um modelo de Rede Neural MLP ou até mesmo utilizando Regressão Logística, cujos resultados no comparativo demonstrou ser mais eficaz.

4.2 Rede Neural MLP

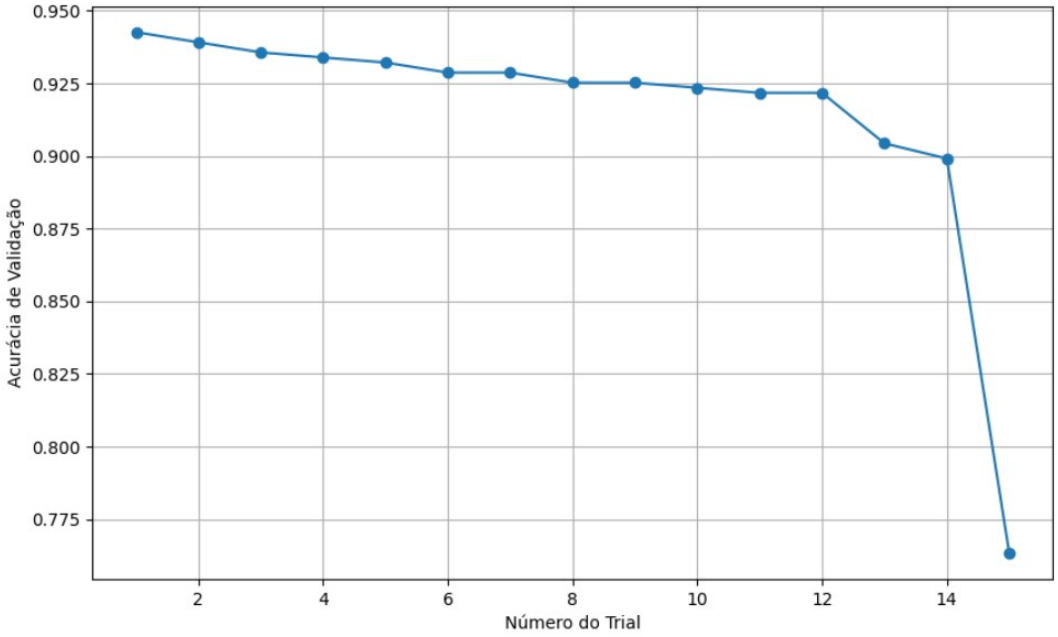
A rede neural MLP foi submetida a um processo de ajuste de hiperparâmetros utilizando o método de busca aleatória, visando otimizar sua performance na tarefa de classificação entre medições fraudulentas e não fraudulentas. Após a busca, o melhor modelo foi selecionado com base na métrica de acurácia no conjunto de validação, conforme detalhe a seguir:

Ajustes de Hiperparâmetros:

- O número de camadas (*num_layers*) variou entre 2 e 20, sendo o melhor resultado obtido com 7 camadas;
- As unidades (*units*) em cada camada variaram entre 32 e 512, com diferentes configurações em cada *trial*;
- A taxa de aprendizado (*learning_rate*) variou entre 0.0001, 0.001 e 0.01, sendo 0.01 a escolhida para o melhor modelo.

O Figura 18 mostra a evolução da acurácia de validação ao longo dos trials de ajuste de hiperparâmetros. Cada ponto no gráfico representa a acurácia de validação do melhor modelo encontrado até o momento em cada trial. O objetivo é observar se a acurácia de validação está aumentando ou estagnando ao longo dos trials, ajudando a entender a eficácia do processo de busca de hiperparâmetros.

Figura 18 – Acurácia de Validação ao Longo dos Trials.



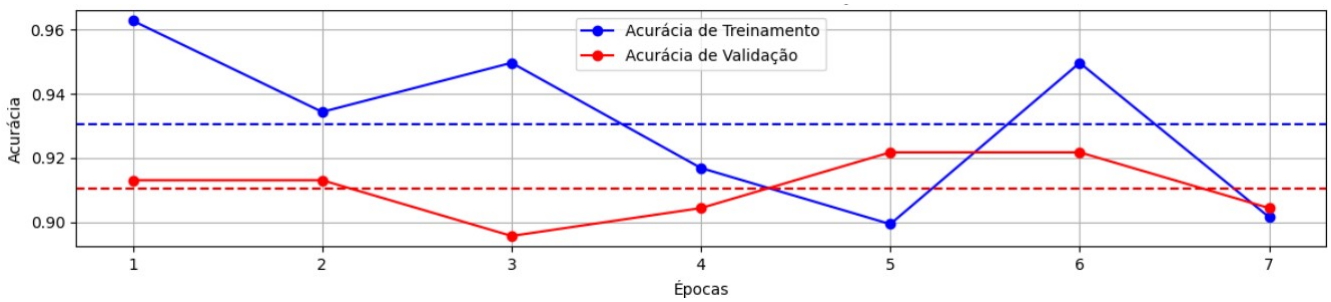
Fonte: Autor (2024).

Melhores Épocas de Treinamento:

- O treinamento da rede neural MLP ocorreu ao longo de várias épocas, com o modelo sendo avaliado periodicamente no conjunto de validação;
- O melhor desempenho foi alcançado após um total de épocas, durante as quais o modelo aprendeu a representação dos dados e ajustou seus pesos para otimizar a classificação.

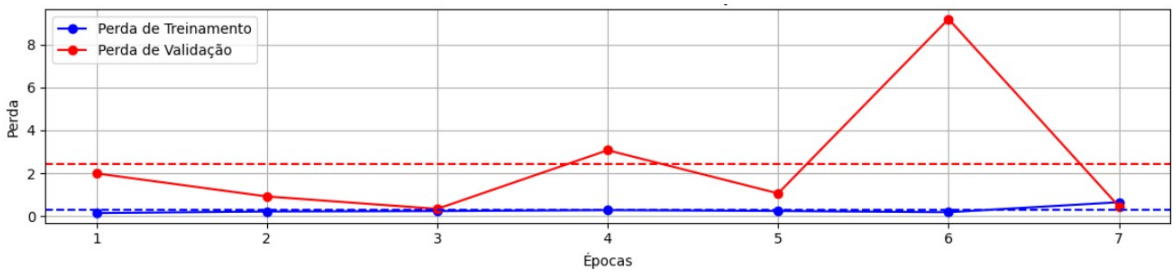
As Figuras 19 e 20 exibem a precisão (acurácia e a perda) do modelo tanto no conjunto de treinamento quanto no conjunto de validação ao longo das épocas de treinamento, demonstrando que o modelo pode fazer previsões corretas evolui à medida que ele aprende com os dados.

Figura 19 – Acurácia de Treinamento e Validação.



Fonte: Autor (2024).

Figura 20 – Perda de Treinamento e Validação.



Fonte: Autor (2024).

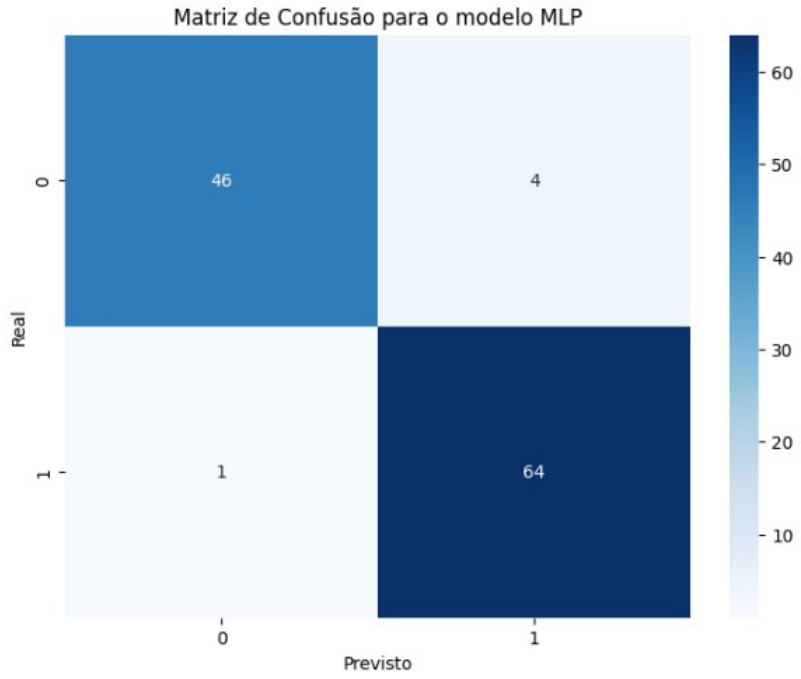
Acurácias do Modelo:

- No conjunto de treino, a rede neural MLP alcançou uma acurácia de aproximadamente 96.06%, indicando um alto grau de precisão na classificação das medições;
- No conjunto de validação, a acurácia foi ligeiramente menor, em torno de 94.78%, mas ainda assim demonstrando uma performance robusta na generalização para dados não vistos durante o treinamento.

Com base nestes resultados é possível afirmar que o modelo de rede neural MLP superou o modelo de regressão linear, indicando que foi capaz de capturar relações mais complexas nos dados. O percentual de erro associado a este modelo foi de 5.22%.

Quando avaliada a matriz de confusão, apresentada na Figura 21, é possível verificar que o modelo de Redes Neurais MLP apresentou resultados muito promissores. Das 115 instâncias, 46 foram corretamente classificadas como negativas e 64 como positivas. O modelo cometeu apenas 5 erros de classificação, sendo 4 instâncias negativas erroneamente classificadas como positivas e 1 instância positiva erroneamente classificada como negativa. Esses resultados indicam uma capacidade significativa do modelo em discriminar entre as classes, com uma taxa muito baixa de erros.

Figura 21 – Matriz de Confusão - Rede Neural MLP.



Fonte: Autor (2024).

4.3 Comparação de Modelos

Adicionalmente, foram aplicados outros métodos à mesma base de dados, cujo comparativo, Tabela 6, oferece uma visão clara dos resultados obtidos por cada modelo, destacando suas respectivas forças e desempenho relativo com base nas métricas selecionadas.

Tabela 6 – Comparação das Metricas de Desempenho dos Modelos.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1 Score**	**AUC**	**Índice Kappa**
Regressão Linear	0.9304	0.9306	0.9304	0.9303	N/A	0.8578
Regressão Logística	0.9478	0.9486	0.9478	0.9479	0.9751	0.8943
Árvore de Decisão	0.9391	0.9394	0.9391	0.9392	0.9392	0.8764
Floresta Aleatória	0.9391	0.9394	0.9391	0.9392	0.9882	0.8764
Perceptron Multicamadas (MLP)	0.9478	0.9478	0.9478	0.9478	0.9708	0.8938

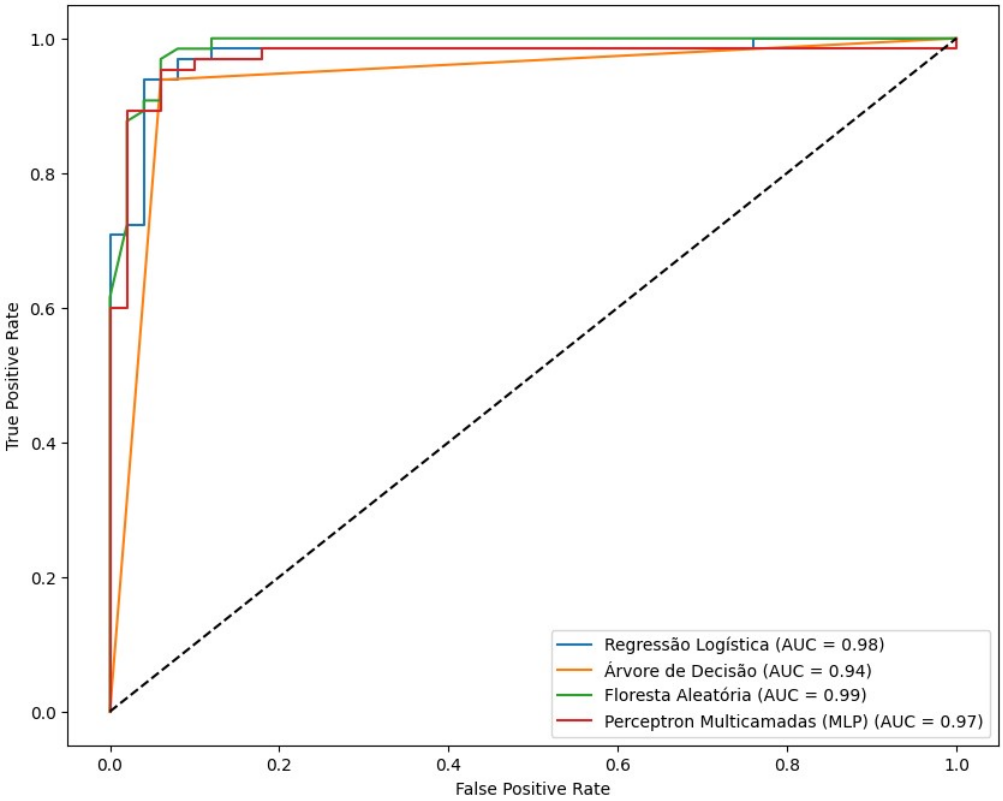
Fonte: Autor (2024).

Os resultados comparativos entre os modelos mostram que a Regressão Logística e Rede Neural (MLP) atingiram a maior acurácia (0.9478), indicando uma excelente capacidade de generalização nos dados de validação. A Floresta Aleatória apresentou um AUC superior (0.9882), sugerindo uma melhor performance em termos de discriminação entre as classes positivas e negativas.

Os valores de precisão, recall e F1 Score são bastante elevados para todos os modelos, com ligeiras variações que podem ser consideradas dentro de uma margem de erro aceitável. O Índice Kappa, que mede a concordância entre as previsões do modelo e as classificações reais corrigindo para o acaso, é também elevado, destacando a robustez dos modelos.

A análise detalhada dos resultados, Figura 22 sugere que, embora a Regressão Logística e a Rede Neural MLP tenham alcançado a maior acurácia, a Floresta Aleatória apresentou o melhor AUC, indicando que este pode ser o modelo mais eficaz em termos de balanceamento entre sensibilidade e especificidade.

Figura 22 – Curvas ROC.



Fonte: Autor (2024).

5 Considerações Finais

Com base nos resultados obtidos deste trabalho, foi possível concluir que a aplicação de redes neurais para a classificação de medições manuais de madeira se mostrou uma abordagem eficaz para identificar padrões complexos e distinguir entre registros legítimos e fraudulentos.

Inicialmente, o objetivo foi otimizar a detecção precoce de práticas fraudulentas no processo de medição de madeira, especialmente nos pátios intermediários, onde a medição manual por régua é mais suscetível a erros e fraudes. Para alcançar esse objetivo, adotou-se uma metodologia que envolveu a seleção e extração cuidadosa dos dados, seguida por etapas de pré-processamento e implementação de modelos de regressão linear e redes neurais.

Os resultados da regressão linear revelaram uma capacidade moderada de explicar a variabilidade nos dados, com uma acurácia de aproximadamente 93.04% no conjunto de validação. Embora a regressão linear tenha apresentado limitações na classificação de instâncias em categorias discretas, os indicadores como o coeficiente de determinação (R^2), erro quadrático médio (MSE) e erro absoluto médio (MAE) forneceram insights valiosos sobre o desempenho do modelo.

Por outro lado, a Rede Neural MLP apresentou resultados mais promissores, com uma acurácia significativamente maior tanto no conjunto de treino (aproximadamente 96.06%) quanto no conjunto de validação (aproximadamente 94.78%). O processo de ajuste de hiperparâmetros permitiu a seleção do melhor modelo, demonstrando a capacidade da rede neural em capturar relações complexas nos dados e fornecer uma classificação precisa das medições.

Ao avaliar a matriz de confusão da Rede Neural MLP, observa-se resultados muito promissores, com uma taxa muito baixa de erros de classificação. Dos 115 registros avaliados, apenas 5 foram classificados de forma incorreta, indicando uma capacidade significativa do modelo em discriminar entre as classes de fraude e não fraude.

Portanto, a aplicação de redes neurais para a detecção de fraudes no processo de medição de madeira apresenta vantagens significativas sobre abordagens tradicionais, como a regressão linear.

Para trabalhos futuros, a expansão da análise com a inclusão de novas variáveis, bem como a identificação de outros métodos de aprendizado de máquina, como o comparativo de desempenho realizado ao final do trabalho, de forma a aprimorar o desempenho do modelo de classificação de fraudes.

Referências

AZEVEDO, T. R.; LEITE, N. B. O amadurecimento do fomento florestal. **Revista Opiniões2**, 2024. 5

BERTOLA, V. **Modelling and Experimentation in Two-Phase Flow**. [S.l.]: Springer Vienna, 2003. ISBN 978-3-211-20757-4. 7, 9

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, American Association for Artificial Intelligence, v. 17, n. 3, p. 37–54, 1996. 2

GERON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. [S.l.]: O'Reilly Media, 2019. 20

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. 14, 15

HÄGGLUND, B. Measurement strategies for stacked timber. **Holz als Roh-und Werkstoff**, Springer, v. 64, n. 6, p. 430–437, 2006. 11

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. [S.l.]: Elsevier, 2011. 17

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. [S.l.]: Springer, 2009. 14

HEATON, J. **Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks**. [S.l.]: Heaton Research, Inc., 2018. 15

IBA. **Benefícios sociais**. 2019. Acesso em 25 de abril de 2024. Disponível em: <https://www.iba.org/dados-estatisticos>. 1

KEEPERS, J. C. **Manejo de Florestas e o Efeito Gaiola**. São Paulo, Brasil: Editora Florestal, 1945. 9

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014. 15

MACHADO, C. C. **Colheita Florestal**. Viçosa, MG: UFV, 2002. 5

MACHADO, C. C.; SILVA, J. R. M. da; LIMA, J. T.; SOUZA, A. D. de. Estimativa do volume de madeira em pilhas. **Revista Árvore**, Viçosa, v. 27, n. 5, p. 657–664, 2003. 9

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. [S.l.]: Wiley, 2012. 13

NIELSEN, M. A. **Neural Networks and Deep Learning**. [S.l.]: Determination Press, 2015. 14

QUANTITATIVOS, C. C. de M. Metrvm - revista brasileira de estatística e econometria. **Metrvm**, v. 2, n. 2, 2002. Acesso em 25 de abril de 2024. Disponível em: <http://cmq.esalq.usp.br/wiki/lib/exe/fetch.php?media=publico:metrvm:metrvm-2002-n02.pdf>. 6, 8

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, Nature Publishing Group, v. 323, n. 6088, p. 533–536, 1986. 14

SIQUEIRA, J. O.; CANTO, A. d. C.; MORAES, L. F. d. Fomento florestal: benefício para empresas e produtores. In: **Anais do II Seminário Internacional sobre Manejo Florestal**. Curitiba: IPEF, 2004. p. 1–8. Acesso em 29 de abril de 2024. Disponível em: <http://www.ipef.br/eventos/iisimf/docs/trabalhos/siqueira.pdf>. 4

STIGLER, S. M. **The History of Statistics: The Measurement of Uncertainty before 1900**. [S.l.]: Harvard University Press, 1986. 13

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. 14

TORQUATO, J. R. Estudo da variação do fator de empilhamento. **Revista de Engenharia Florestal**, v. 2, n. 1, p. 23–29, 1983. 9

WOODTECH MEASUREMENT. **Logmeter**. 2014. Acesso em 7 de maio de 2024. Disponível em: <https://www.woodtechms.com/_files/ugd/b374e6_751359a46f824ff5a23682ae9206fe0c.pdf>. 10, 11

WOODTECH, P. **Quantificação da Madeira Recebida**. 2007. Acesso em 9 de maio de 2024. Disponível em: <https://www.eucalyptus.com.br/artigos/2007_Quantificazao+Madeira_Recebida.pdf>. 7