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

Utilização de Redes Neurais Artificiais e Técnica SHAP para Predição do Teor de Sólidos Totais durante a Secagem de Lodo Biológico de uma Fábrica de Celulose Kraft

Rafael Gonçaves Silva

João Monlevade, MG
2024

Rafael Gonçaves Silva

Utilização de Redes Neurais Artificiais e Técnica SHAP para Predição do Teor de Sólidos Totais durante a Secagem de Lodo Biológico de uma Fábrica de Celulose Kraft

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. Thiago Augusto de Oliveira Silva

Coorientador: MSc. Ronaldo Neves Ribeiro

João Monlevade, MG

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S586u Silva, Rafael Gonçalves.
Utilização de redes neurais artificiais e técnica shap para predição do teor de sólidos totais durante a secagem de lodo biológico de uma fábrica de celulose kraft. [manuscrito] / Rafael Gonçalves Silva. - 2024.
44 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Thiago Augusto de Oliveira Silva.
Produção Científica (Especialização). Universidade Federal de Ouro Preto. Departamento de Engenharia de Produção.

1. Aprendizado do computador. 2. Controle preditivo. 3. Estatística matemática. 4. Industria de celulose. 5. Lodo - Secagem. 6. Processo decisório. 7. Redes neurais (Computação). I. Silva, Thiago Augusto de Oliveira. II. Universidade Federal de Ouro Preto. III. Título.

CDU 519.2:004.8

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

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

FOLHA DE APROVAÇÃO

Rafael Gonçalves Silva

Utilização de Modelagem Preditiva e Análise de Dados no Processo de Secagem de Lodo Biológico em Uma Fábrica de Celulose Kraft

Trabalho de conclusão de curso apresentado ao curso de Especialização 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 09 de maio de 2024

Membros da banca

Dr. Thiago Augusto de Oliveira Silva - Orientador - Universidade Federal de Ouro Preto
Dr. Carlos Henrique Gomes Ferreira - Universidade Federal de Ouro Preto
Me. Ronaldo Neves Ribeiro - Celulose Nipo Brasileira - CENIBRA

Thiago Augusto de Oliveira Silva, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 29/07/2024

Documento assinado eletronicamente por **Thiago Augusto de Oliveira Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 06/08/2024, às 12:54, 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 **0750943** e o código CRC **80CF32EB**.

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

Dedico este trabalho a toda minha família

Agradecimentos

Agradeço a Deus por guiar sempre os meus passos e pela força concedida para concretizar este desafio.

À minha esposa Heloana, pelo carinho e companheirismo.

Ao meu filho Davi, que nasceu durante o curso, sendo um recém-nascido nos momentos mais desafiadores desta jornada. Embora ainda muito pequeno para compreender a minha busca por um futuro melhor, agradeço por sua presença, que trouxe luz e inspiração aos dias mais difíceis. Davi, você é a razão de todos os esforços, e eu espero ser um exemplo inspirador para o seu próprio caminho à medida que você cresce.

Ao meu avô, Geraldo Magela (in memoriam), pelo legado de educação e por ser o meu maior exemplo de vida.

À minha mãe, Mariza, pelo amor incondicional e dedicação.

A todos os colegas do curso que generosamente compartilharam seus conhecimentos.

Ao professor e orientador Dr. Thiago Silva, agradeço por não medir esforços para me auxiliar nesta última etapa do curso.

Meus sinceros agradecimentos.

“A persistência é o menor caminho do êxito. Charles Chaplin”

Resumo

Este trabalho investiga a baixa eficiência de um processo de secagem de lodo biológico em uma fábrica de celulose Kraft, um problema crucial devido ao impacto significativo na eficiência operacional e sustentabilidade ambiental. A pesquisa visa aprimorar o entendimento e a previsibilidade dos resultados desse processo através do desenvolvimento de um modelo preditivo baseado em *Redes neurais artificiais* (RNA), especificamente *MultiLayer perceptron* (MLP), e utilizando a metodologia de análise de dados *Cross industry standard process for data mining* (CRISP-DM). Para realizar a modelagem preditiva, características relevantes foram analisadas utilizando a técnica de *SHapley additive exPlanations* (SHAP), que proporciona uma interpretação detalhada dos fatores que mais influenciam as previsões. O ajuste fino dos hiperparâmetros foi realizado com o auxílio da ferramenta *Keras-Tuner*, permitindo a exploração de diferentes configurações da rede neural para melhorar a precisão do modelo. Os resultados obtidos indicam que o modelo desenvolvido conseguiu prever com alta precisão a consistência dos sólidos após a secagem, com um *Mean absolute error* (MAE) de 0,84% e um coeficiente de determinação R^2 de 0.78, confirmando que o modelo é capaz de capturar as complexidades do processo de secagem, oferecendo uma ferramenta robusta e confiável para a tomada de decisão na gestão de resíduos. O estudo reforça a aplicabilidade das técnicas de aprendizado de máquina na indústria de celulose, sugerindo que tais abordagens podem proporcionar *insights* valiosos para um melhor controle de operações industriais, com implicações diretas na eficiência e sustentabilidade. Recomenda-se a validação futura do modelo em outras instalações industriais para explorar sua adaptabilidade e eficácia em diferentes contextos operacionais.

Palavras-chave: secagem de lodo biológico, redes neurais MLP, celulose Kraft, aprendizado supervisionado, modelagem preditiva, SHAP.

Abstract

This work investigates the low efficiency of a biological sludge drying process in a Kraft pulp factory, a critical issue due to its significant impact on operational efficiency and environmental sustainability. The research aims to enhance the understanding and predictability of the process outcomes through the development of a predictive model based on *Artificial neural networks* (ANN), specifically MLP, using the CRISP-DM data analysis methodology. To perform the predictive modeling, relevant features were analyzed using the SHAP technique, which provides a detailed interpretation of the factors most influencing the predictions. Hyperparameter tuning was conducted with the assistance of the Keras-Tuner tool, enabling the exploration of different neural network configurations to improve the model's accuracy. The results obtained indicate that the developed model was able to predict with high precision the consistency of the solids after drying, with a MAE of 0.84% and a Coefficient of Determination R^2 of 0.78, confirming that the model is capable of capturing the complexities of the drying process, providing a robust and reliable tool for decision-making in waste management. The study reinforces the applicability of machine learning techniques in the pulp industry, suggesting that such approaches can provide valuable insights for better control of industrial operations, with direct implications for efficiency and sustainability. Future validation of the model in other industrial facilities is recommended to explore its adaptability and effectiveness in different operational contexts.

Keywords: biological sludge drying, MLP, Kraft pulp, supervised learning, predictive modeling, SHAP.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AED Análise exploratória de dados

AM Aprendizado de máquina

ANN *Artificial neural networks*

CONAMA *Conselho nacional de meio ambiente*

CRISP-DM *Cross industry standard process for data mining*

IA Inteligência artificial

KDD *Knowledge discovery in databases*

KDE *Kernel density estimation*

LSTM *Short-term memory*

MAE *Mean absolute error*

MLP *MultiLayer perceptron*

MSE *Mean squared error*

OPC *OLE for process control*

PIMS *Plant information management system*

PNRS *Política nacional de resíduos sólidos*

Q-Q *Quantil-Quantil*

RMSE *Root mean squared error*

RNA *Redes neurais artificiais*

SDCD Sistema digital de controle distribuído

SHAP *SHapley additive exPlanations*

TCC Trabalho de conclusão de curso

TS *Sólidos Totais*

Lista de símbolos

b	Viés
f	Função de ativação
μ	Média dos dados
N	Número total de observações
R^2	Coeficiente de determinação
σ	Desvio padrão dos dados
w_i	Peso associado à entrada i
x	Valor original do dado
x_i	Entrada i
Y	Saída do neurônio
$\bar{y}$	Média dos valores observados
$\hat{y}_i$	Valores preditos pelo modelo
y_i	Valores reais

Sumário

1 Introdução

A indústria de celulose é um setor chave da economia mundial, com o Brasil alcançando recordes na produção de celulose, destacando-se com 25 milhões de toneladas em 2022 (IBÁ, 2023). Um dos desafios enfrentados por esse setor é o gerenciamento eficiente do lodo biológico secundário, um subproduto do tratamento de efluentes nas fábricas de celulose kraft branqueada. Este lodo, caracterizado por alta carga orgânica e complexidade físico-química, demanda processos otimizados de secagem para sua adequada disposição ou reciclagem (NAVAEE-ARDEH; BERTRAND; STUART, 2006).

O processo de secagem do lodo biológico é crucial para reduzir seu volume e facilitar sua reciclagem ou disposição final, sendo intensivo em energia e sujeito a variações de desempenho devido a um leque de variáveis operacionais e características do lodo (FAHIM *et al.*, 2019). A otimização dessa etapa é essencial, prometendo melhorias na eficiência operacional, redução de custos e avanços significativos em sustentabilidade e conformidade regulatória.

Neste contexto, a ciência de dados oferece uma gama de ferramentas e técnicas que podem ser empregadas para entender, modelar e prever o comportamento do processo de secagem de lodo biológico. Particularmente, as técnicas de aprendizado de máquina, incluindo redes neurais MLP, apresentam um potencial significativo para capturar as complexidades e dinâmicas não-lineares do processo. Através da modelagem preditiva e análise de dados, é possível identificar variáveis-chave que influenciam a eficiência da secagem, prever o desempenho do processo sob diferentes condições operacionais, e propor ajustes e otimizações que conduzam a melhorias tangíveis.

Para o desenvolvimento do modelo preditivo, diversas variáveis-chave foram identificadas como influentes. Essas variáveis foram selecionadas com base em sua relevância e impacto potencial no processo de secagem. As principais variáveis incluídas no estudo foram a temperatura do ar fresco (*TC56_PV*), umidade relativa do ar de exaustão (*AI57_PV*), consistência de entrada do lodo (*AI60_PV*), taxa de alimentação real do secador (*21SF004_TAL*), temperaturas segmentadas do ar (*TC60_PV* a *TC89_PV*) e velocidade das esteiras (*SC46_PV*, *SC47_PV*). Essas variáveis foram analisadas utilizando técnicas de aprendizado de máquina e a metodologia SHAP para entender sua contribuição e impacto no modelo preditivo. A análise detalhada dessas variáveis permitiu desenvolver um modelo robusto capaz de prever a consistência de saída do lodo com alta precisão, contribuindo para a otimização do processo e melhoria da eficiência operacional.

- **Hipótese primária (H1)**: A aplicação de técnicas de modelagem preditiva e análise de dados proporcionará novas perspectivas científicas sobre o processo de secagem de lodo biológico em fábricas de celulose Kraft, enriquecendo a base de conhecimento existente na área.
- **Hipótese secundária (H2)**: Variáveis como temperaturas de ar do secador, consistência de entrada, umidade relativa da saída do secador, pressões de ar quente, taxa de alimentação real do secador, e velocidades das esteiras superior e inferior têm um impacto significativo na eficiência e eficácia do processo de secagem de lodo biológico.

1.1 Objetivo geral

Analisar o processo de secagem de lodo biológico em uma fábrica de celulose Kraft branqueada utilizando técnicas de modelagem preditiva e análise de dados, com o intuito de aprimorar a eficiência operacional e promover uma gestão mais sustentável dos resíduos industriais.

1.1.1 Objetivos específicos

- Identificar as principais variáveis que influenciam a consistência de saída do lodo biológico e seu impacto no processo de secagem;
- Desenvolver e implementar um modelo de aprendizado de máquina para prever a consistência de saída do lodo durante a secagem, visando melhorar a precisão do controle do processo;
- Avaliar a eficácia do modelo preditivo em termos de melhoria na eficiência operacional e redução de custos associados ao tratamento e descarte de lodo;
- Propor recomendações baseadas nos resultados obtidos para otimizar o processo de secagem e minimizar o impacto ambiental.

1.2 Contribuições

Este trabalho visa contribuir para a literatura científica na interseção da ciência de dados e gestão de resíduos industriais, fornecendo *insights* e metodologias aplicáveis para a otimização do processo de secagem de lodo biológico. As descobertas deste estudo têm o potencial de beneficiar a indústria de celulose, oferecendo caminhos para a adoção de tecnologias de análise de dados avançadas que promovam operações mais eficientes, econômicas e ambientalmente responsáveis.

1.3 Organização do trabalho

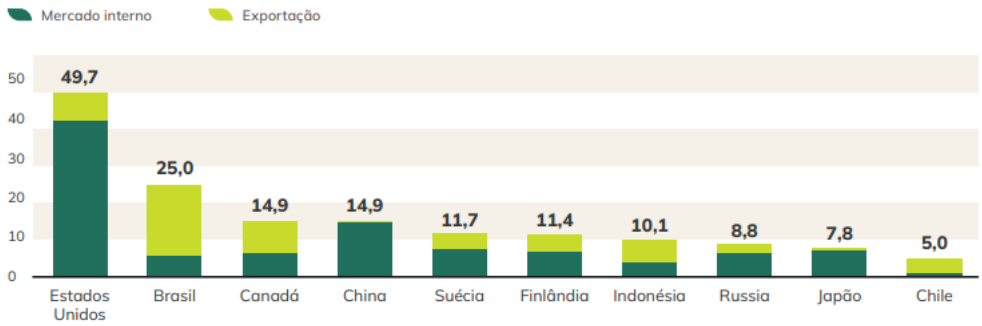
Este trabalho está organizado em cinco capítulos para orientar o leitor através do estudo de forma clara e estruturada. O primeiro capítulo introduz o tema e define os objetivos da pesquisa. O segundo capítulo aborda a revisão da literatura, fornecendo uma base teórica sólida. No terceiro capítulo, descreve-se a metodologia empregada para coletar e analisar os dados. Os resultados são discutidos no quarto capítulo, onde as principais descobertas são destacadas. Finalmente, o quinto capítulo conclui o trabalho, resumindo os achados e sugerindo futuras direções para pesquisa. Apêndices e referências bibliográficas complementam o estudo.

2 Revisão da Literatura

2.1 Contexto da indústria de celulose no Brasil

A indústria de celulose e papel, um dos mais expressivos segmentos da economia brasileira, tem sua importância reiterada no cenário das exportações nacionais. Em um ano de recordes, 2022 presenciou o Brasil atingindo a marca de 25 milhões de toneladas em produção de celulose. O desempenho nas exportações foi igualmente impressionante, com o país exportando 19,1 milhões de toneladas e gerando receitas de exportação que alcançaram US$ 14,29 bilhões. Esses números reafirmam o Brasil como o principal exportador global de celulose, com uma concentração significativa de suas atividades nas regiões Sul e Sudeste. A Figura 1 ilustra a posição do Brasil como o segundo maior produtor mundial de celulose em 2021, apenas atrás dos Estados Unidos, ressaltando a sua influência substancial no mercado global (IBÁ, 2023).

Figura 1 – Principais países produtores de celulose em 2021 (milhões de toneladas)



Fonte: (IBÁ, 2023)

Além do seu impacto econômico, o setor é também um campo de grande amplitude tecnológica e de inovação. A transformação de recursos vegetais em polpa e, por fim, em papel, está intrinsecamente ligada à disponibilidade e qualidade dos recursos naturais e ao investimento contínuo em inovações voltadas para a sustentabilidade, fortalecendo a competitividade brasileira no cenário internacional (PINHEIRO, 2008). Essa interação dinâmica entre economia e meio ambiente é ilustrada pelas práticas de gestão de resíduos da indústria, especialmente na forma como trata seus efluentes e subprodutos.

Os lodos de tratamento de efluentes, categorizados como não perigosos mas não inertes, representam um desafio particular para a indústria. A composição variável desses resíduos, influenciada pelo tipo de papel produzido e pelos processos químicos envolvidos, impõe complexidades adicionais às estratégias de manejo e reciclagem. A indústria tem, portanto, o incentivo para buscar e implementar tecnologias que possam reaproveitar esses resíduos, uma tendência que reflete uma mudança global em direção a práticas mais sustentáveis (TOCZYŁOWSKA-MAMIńSKA, 2017).

A adoção dessas tecnologias inovadoras é fundamental, não apenas como uma resposta às exigências regulatórias e ao imperativo ambiental, mas também como uma estratégia para melhorar a eficiência operacional e reduzir custos. O lodo biológico, em particular, com sua rica composição orgânica, exige métodos de secagem eficientes e sustentáveis. A otimização desses processos é crucial e se apresenta como um tema de estudo emergente, focado na redução do impacto ambiental, na melhoria da qualidade do produto final e na viabilidade econômica da indústria (COSTA; CRISTELLI; PATROCINIO, 2021).

2.2 Processos de tratamento de efluentes e geração de lodo

Os processos de tratamento de efluentes são essenciais nas operações das fábricas de celulose, destacando-se pelos desafios técnicos e impactos ambientais envolvidos. Esses processos visam não apenas à conformidade com normativas ambientais rigorosas, mas também ao aprimoramento da sustentabilidade operacional. Como destacado por Metzner (2018), o tratamento de efluentes é crucial para reduzir os contaminantes provenientes da fabricação de celulose, garantindo que os padrões de qualidade da água sejam mantidos.

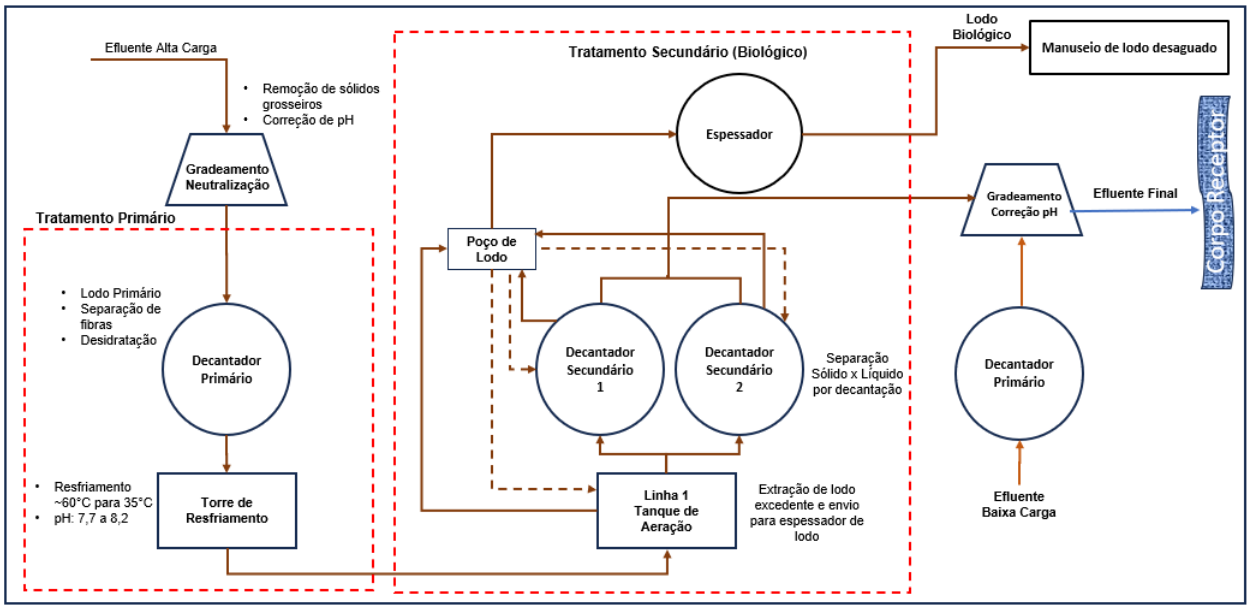
O tratamento inicia-se com o primário, onde sólidos grosseiros são removidos e o pH é ajustado. Segue-se o secundário, que utiliza processos biológicos para decompor matéria orgânica, com microrganismos em tanques de aeração e decantadores secundários para aumentar a eficiência da degradação e separação da biomassa do efluente tratado (SEADI; RUTZ, 2019).

Durante estes processos, o lodo biológico é produzido como subproduto. Este lodo é uma mistura complexa de materiais orgânicos e inorgânicos, cuja composição varia significativamente dependendo da matéria-prima e dos processos químicos empregados (LIMEIRA, 2016). As propriedades físico-químicas do lodo, como pH e teor de umidade, são cruciais para a seleção de tratamentos e reciclagem. O processo de secagem do lodo é essencial para reduzir seu volume e peso, facilitando o manejo e disposição final, como apontam (SILVA; ROSA, 2016). De acordo com um estudo recente, as fábricas de celulose Kraft novas e modernas podem gerar aproximadamente 30 a 40 kg de lodo seco por tonelada de celulose produzida (COSTA; CRISTELLI; PATROCINIO, 2021).

Nesse contexto, a *Política nacional de resíduos sólidos* (PNRS) (Lei Nº 12.305/2010) desempenha um papel fundamental ao estabelecer diretrizes para a gestão integrada e o gerenciamento ambientalmente adequado dos resíduos sólidos, incluindo lodos de processos industriais. A PNRS incentiva a minimização da geração de resíduos por meio de práticas sustentáveis de reutilização, reciclagem e conversão de resíduos em energia, alinhando as operações das fábricas de celulose Kraft com os objetivos de sustentabilidade nacional (BRASIL, 2010).

Na Figura 2, é mostrado o fluxograma do processo de tratamento de efluentes e geração de lodo, ilustrando claramente as etapas descritas:

Figura 2 – Processo de tratamento de efluentes de uma fábrica de celulose kraft



Fonte: elaborado pelo autor

Técnicas avançadas, como digestão anaeróbica, compostagem e incineração, são exploradas para adicionar valor ao lodo e diminuir seu impacto ambiental. A Resolução *Conselho nacional de meio ambiente* (CONAMA) nº 430/2010 estabelece padrões rigorosos para o tratamento e disposição do lodo, garantindo práticas que minimizem os impactos ambientais (AMBIENTE, 2010). A gestão eficiente deste resíduo é crucial para garantir a sustentabilidade ambiental e a viabilidade econômica da indústria de celulose.

2.3 Tecnologias de secagem de lodo biológico

A secagem de lodo biológico é um estágio crítico no tratamento de resíduos em fábricas de celulose Kraft, sendo vital para reduzir o teor de umidade do lodo, permitindo sua disposição adequada ou utilização energética. Segundo (EIKELBOOM *et al.*, 2018), a escolha do método de secagem influencia diretamente a eficiência energética, os impactos ambientais e a qualidade do produto seco. As tecnologias de secagem variam desde métodos tradicionais, como secagem térmica, até abordagens inovadoras, como secagem solar e por micro-ondas.

No tratamento convencional, o lodo é exposto a um ambiente controlado, onde o calor aplicado de forma indireta evita a degradação do material orgânico. A eficácia desse processo depende da eficiência do intercâmbio térmico e da capacidade do sistema em reutilizar o calor residual, frequentemente oriundo de gases de exaustão de outras operações da planta, como caldeiras de biomassa (SANTOS *et al.*, 2021). A água quente gerada é recirculada através de trocadores de calor, otimizando o perfil térmico necessário para a secagem.

Dentre as tecnologias, os secadores de correia são destacados pela sua eficiência energética e baixo custo operacional, sendo versáteis e adaptáveis a diferentes tipos de lodo. A figura 3 ilustra um exemplo de secador de correia utilizado no processo. A modelagem e simulação são essenciais para otimizar esses sistemas, considerando variáveis como temperatura, umidade relativa e velocidade do ar (HUILIñIR; ALLEN; TRAN, 2017). A escolha do método de secagem e a operacionalização eficiente são cruciais para maximizar a recuperação de energia e minimizar os custos e impactos ambientais.

Figura 3 – Secador de Correia de Média Temperatura

Fonte: (Huber Technology, 2023).

Este panorama ressalta a importância da inovação e da pesquisa contínua no desenvolvimento de métodos de secagem que sejam eficientes e sustentáveis, contribuindo para a gestão ambiental responsável na indústria de celulose.

2.3.1 Sensores de consistência no processo de secagem de lodo biológico

Os sensores de consistência, que medem o teor de *Sólidos Totais* (TS), desempenham um papel crítico no processo de secagem de lodo biológico, especialmente na indústria de celulose, onde a eficiência e a sustentabilidade dos processos de tratamento são de suma importância. A função principal desses sensores é monitorar e controlar o teor de sólidos no lodo ao longo do processo de secagem, fornecendo dados essenciais para a otimização das condições operacionais e garantindo a qualidade do produto final.

Os sensores de TS medem continuamente o teor de sólidos do lodo, permitindo ajustes em tempo real nas variáveis do processo de secagem, como temperatura, pressão e tempo de exposição. Esses ajustes são fundamentais para evitar a secagem excessiva ou insuficiente do lodo, o que pode afetar negativamente tanto a eficiência energética do processo quanto a qualidade do lodo seco. Por exemplo, um teor de sólidos mais alto no produto final significa uma menor quantidade de água residual, o que reduz o custo e o impacto ambiental associados ao transporte e à disposição do lodo. Por outro lado, a secagem excessiva pode levar a um consumo desnecessário de energia e à deterioração das propriedades físico-químicas do lodo, comprometendo seu potencial de reutilização.

Portanto, os sensores de TS são ferramentas indispensáveis no processo de secagem de lodo biológico, facilitando a implementação de práticas de gestão mais eficientes.

2.4 Aprendizado de máquina

O Aprendizado de máquina (AM), um subcampo da Inteligência artificial (IA), concentra-se no desenvolvimento de sistemas que aprendem a partir de dados para realizar previsões ou tomar decisões automaticamente, sem necessidade de programação específica para cada tarefa. Essa área, segundo Mitchell (1997), investiga a construção de algoritmos que podem melhorar seu desempenho com experiências acumuladas.

A essência do aprendizado de máquina reside na capacidade dos sistemas de identificar padrões e fazer inferências com mínima intervenção humana, uma abordagem que, conforme Goodfellow, Bengio e Courville (2016), é aplicável em uma ampla variedade de setores, incluindo saúde, finanças e entretenimento. Esse enfoque transforma teorias computacionais em soluções práticas que aprimoram tanto atividades humanas quanto processos operacionais.

Existem três categorias principais de aprendizado de máquina, cada uma adaptada para abordar diferentes tipos de problemas de dados:

- **Aprendizado Supervisionado:** Segundo Alpaydin (2020), modelos são treinados usando dados rotulados e aprendem a prever resultados futuros.
- **Aprendizado Não Supervisionado:** Como descrito por Hinton e Sejnowski (1999), modelos analisam dados não rotulados para inferir padrões complexos.
- **Aprendizado por Reforço:** Sutton e Barto (2018) explicam que modelos fazem escolhas e aprendem a partir das consequências de suas ações, ajustando suas estratégias para maximizar a recompensa.

Apesar das suas promessas, o aprendizado de máquina enfrenta desafios significativos, como a necessidade de vastas quantidades de dados de alta qualidade. Bishop (2006) destaca que existe também o risco de *overfitting*, onde modelos altamente treinados em conjuntos de dados específicos falham ao serem aplicados a novos dados. Além disso, questões éticas surgem com a coleta intensiva de dados pessoais, levantando preocupações sobre privacidade e a necessidade de desenvolver práticas justas e transparentes, conforme apontado por Shokri e Shmatikov (2015).

2.4.1 Aprendizado supervisionado

O aprendizado supervisionado é uma abordagem em que o modelo é treinado em um conjunto de dados rotulados, aprendendo a mapear entradas para saídas. Esse tipo de aprendizado é essencial para tarefas que requerem previsões precisas, como a classificação de e-mails em *spam* ou não *spam*, ou a previsão de tendências de mercado. O aprendizado supervisionado é caracterizado pelo uso de um conjunto de treinamento, onde cada exemplo de entrada vem com um rótulo correspondente ou resultado desejado, permitindo que o modelo aprenda a relação entre os dois. Após o treinamento, o modelo é testado em dados não vistos para verificar sua capacidade de generalização (MURPHY, 2012).

2.4.2 Redes neurais multicamadas (MLP)

O MLP, ou Redes Neurais Multicamadas, é uma arquitetura de rede neural *feedforward* que consiste em várias camadas de neurônios. Cada neurônio em uma camada está interconectado com todos os neurônios na próxima camada, e o processamento dentro de cada neurônio é realizado através de uma função de ativação não linear. A função que representa a operação de um neurônio individual no MLP é dada por:

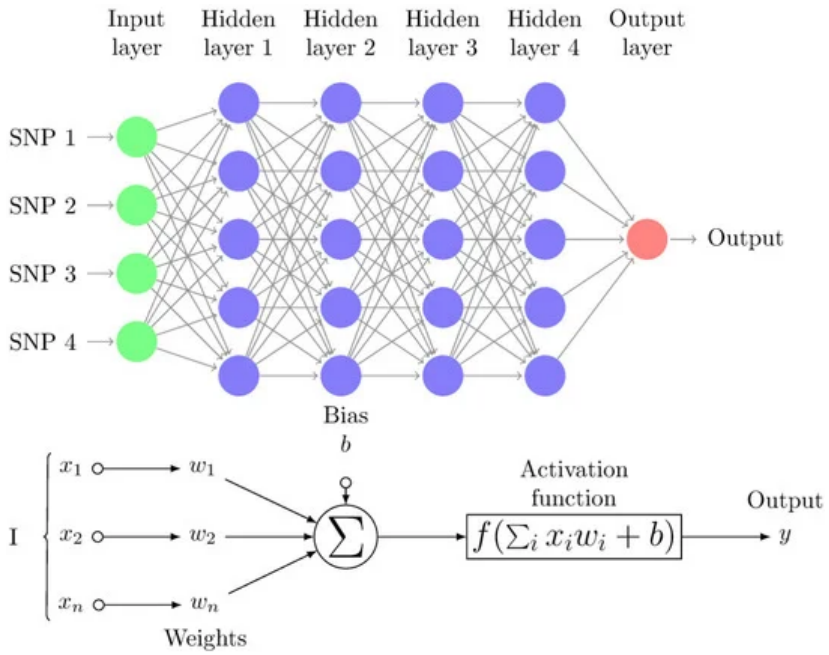
$$Y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.1}$$

onde Y é a saída do neurônio, w_i representa os pesos, x_i são as entradas, b é o viés, e f denota a função de ativação, que introduz não-linearidade ao modelo. O MLP é capaz de modelar complexidades nos dados que modelos mais simples, como a regressão linear, não conseguem, tornando-o adequado para uma ampla gama de tarefas de classificação e regressão.

Segundo Goodfellow, Bengio e Courville (2016), o MLP é um componente fundamental nas aplicações de aprendizado profundo, fornecendo a base para modelos mais complexos usados em tarefas de visão computacional, processamento de linguagem natural e muitas outras áreas de pesquisa em ciência de dados.

Na Figura 4, a arquitetura do MLP é apresentada, detalhando as camadas de neurônios e suas interconexões.

Figura 4 – Arquitetura MLP



Fonte: Pérez-Enciso e Zingaretti (2019, p. 4).

Essa capacidade de capturar relações não lineares e complexas faz do MLP uma ferramenta poderosa na modelagem preditiva e na análise de dados.

2.5 Metodologia SHAP

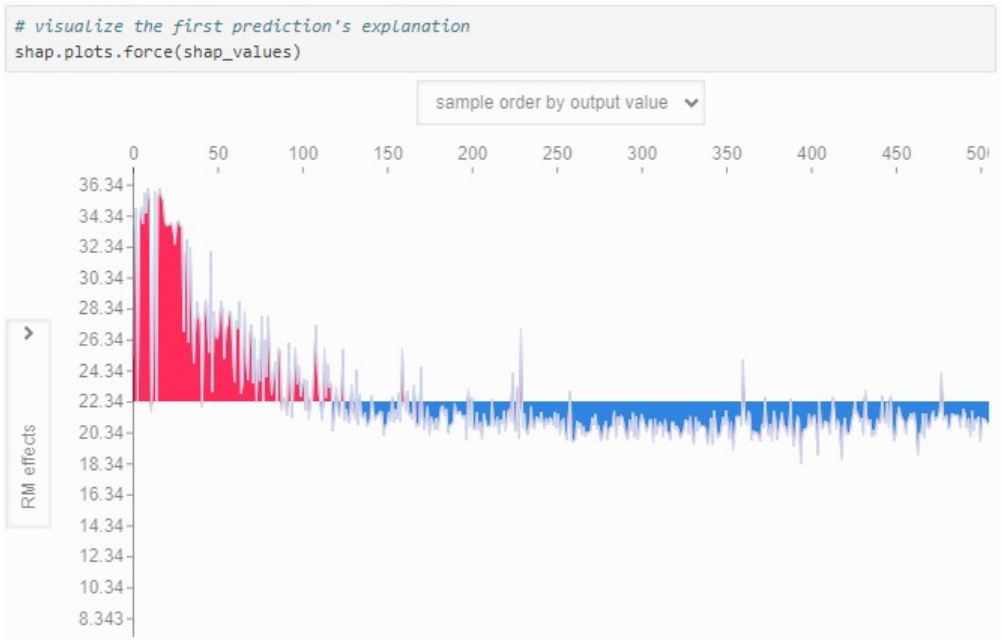
A metodologia SHAP é uma abordagem inovadora e amplamente reconhecida para a interpretação de modelos de aprendizado de máquina. Baseada na teoria dos jogos cooperativos, ela atribui valores a cada característica de entrada de um modelo, quantificando sua contribuição individual para a previsão. Essa técnica é especialmente valiosa para entender modelos complexos, como as redes neurais multicamadas, onde as interações não lineares entre as características podem dificultar a interpretação das previsões.

Segundo Lundberg e Lee (2017), esta metodologia combina os princípios das soluções de *Shapley*, provenientes da teoria dos jogos, com explicações aditivas, oferecendo uma interpretação consistente e localmente precisa das previsões dos modelos de aprendizado de máquina. As soluções de *Shapley* garantem uma distribuição justa do valor total gerado por uma coalizão de características, assegurando que a contribuição de cada uma seja considerada de maneira equitativa.

A aplicação da metodologia envolve os seguintes passos:

- **Treinamento do Modelo**: Inicialmente, um modelo de aprendizado de máquina, como o MLP, é treinado utilizando um conjunto de dados representativo. Este modelo deve ser capaz de capturar as complexidades e interações presentes nos dados.
- **Cálculo dos Valores SHAP**: Utilizando o modelo treinado, os valores são calculados para cada observação no conjunto de dados. Estes valores representam a contribuição marginal de cada característica para a previsão do modelo, considerando todas as possíveis combinações de características.
- **Visualização e Interpretação**: Os valores calculados podem ser visualizados através de diversos tipos de gráficos, como gráficos de barras, gráficos de dependência e gráficos de força. Essas visualizações ajudam a interpretar como cada característica influencia as previsões do modelo, tanto individualmente quanto em interação com outras características.

Figura 5 – Exemplo de Gráfico de Força SHAP



Fonte: (LUNDBERG; LEE, 2021).

A utilização da metodologia SHAP proporciona várias vantagens:

- **Interpretação Consistente**: Os valores garantem uma interpretação consistente das previsões, atribuindo valores justos a cada característica com base em sua contribuição marginal.
- **Transparência do Modelo**: Ao decompor a previsão do modelo em contribuições individuais de cada característica, a metodologia aumenta a transparência, permitindo uma compreensão mais profunda das decisões do modelo.
- **Identificação de Características Importantes**: Facilita a identificação das características mais influentes nas previsões, auxiliando na seleção de características e na otimização de modelos.
- **Análise de Interações**: Permite a análise das interações entre as características, destacando como combinações específicas de características afetam as previsões.

A aplicação da metodologia é especialmente relevante para a análise de processos industriais complexos, como a secagem de lodo biológico em fábricas de celulose. Neste contexto, sua utilização pode revelar *insights* valiosos sobre os fatores que mais influenciam a eficiência do processo de secagem, orientando ajustes operacionais e melhorias de desempenho.

Essa metodologia fundamenta tanto a seção de metodologia quanto a de resultados deste estudo, proporcionando uma base robusta para a interpretação das previsões do modelo MLP e para a análise das contribuições individuais das características do processo.

2.6 Trabalhos relacionados

Este trabalho visa expandir as abordagens de aprendizado de máquina aplicadas ao gerenciamento de resíduos industriais, inspirando-se no estudo de Eric Bröndum, que modelou o processo de espessamento de lodo primário em uma estação de tratamento de águas residuais (BRÖNDUM, 2022). Bröndum utilizou modelos de redes neurais, incluindo *Short-term memory* (LSTM) e MLP, para prever o conteúdo sólido do lodo e otimizar a dosagem de polímeros, uma estratégia essencial para a eficiência do processo e redução de custos.

Da mesma forma, no contexto das fábricas de celulose Kraft branqueada, este trabalho investiga o uso de técnicas de modelagem preditiva, particularmente o MLP, para entender o comportamento do processo de secagem de lodo biológico. Através da modelagem e análise de dados, procura-se identificar variáveis-chave que influenciam a eficiência da secagem, conforme as hipóteses do estudo, sem, no entanto, chegar ao nível de otimização do processo neste estágio.

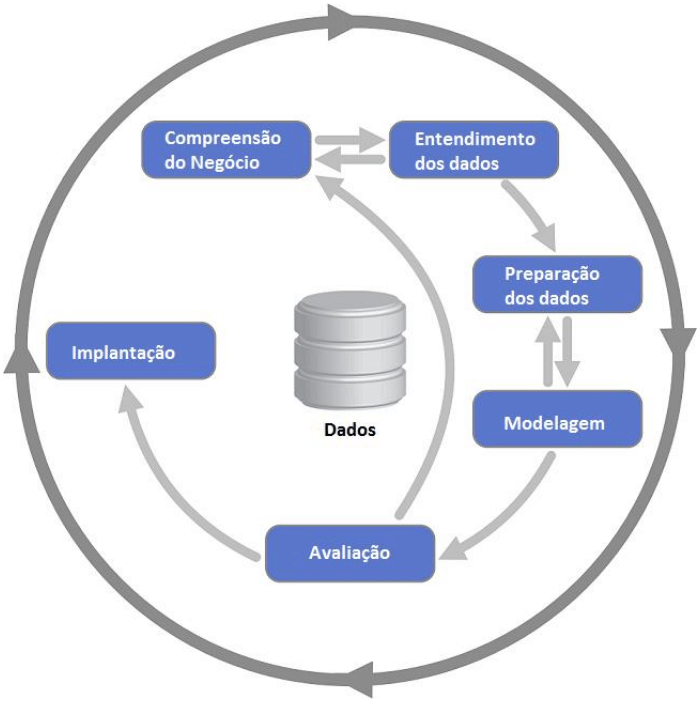
Os *insights* do estudo de Bröndum são particularmente valiosos para este trabalho, pois demonstram como a modelagem preditiva pode ser utilizada para controlar e otimizar processos industriais complexos. Esta abordagem é explorada neste Trabalho de conclusão de curso (TCC) para entender melhor o processo de secagem de lodo, enfatizando a relevância das técnicas de aprendizado de máquina para fornecer novas perspectivas sobre processos industriais (BRÖNDUM, 2022).

3 Metodologia

Neste estudo, adotou-se a metodologia CRISP-DM para definir os processos de ajuste do conjunto de dados e predizer a consistência do lodo na saída do secador de correia. Reconhecida por sua abordagem estruturada e iterativa, esta metodologia é ideal para lidar com os desafios específicos deste projeto e é preferida por 42% dos profissionais de *data mining*, segundo KDnuggets (2007). A preferência pelo CRISP-DM em relação ao processo *Knowledge discovery in databases* (KDD) decorre de sua abordagem mais centrada na resolução de problemas de negócios e maior flexibilidade para adaptar-se a diferentes fases operacionais.

Este modelo inclui seis fases essenciais: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e, tipicamente, implantação. Segundo Wirth e Hipp (2000), o CRISP-DM proporciona uma "estrutura clara e iterativa que facilita a adaptação às necessidades específicas de diferentes indústrias e processos". Clifton e Thuraisingham (2001) também destacam a flexibilidade destas fases, que são particularmente adaptáveis a diversos cenários. A figura abaixo ilustra o fluxograma da metodologia, que visualmente representa essas fases e sua sequência lógica.

Figura 6 – Fluxograma da Metodologia CRISP-DM



Fonte: (RAMOS *et al.*, 2020).

Segue abaixo uma breve explicação de cada fase da metodologia:

- **Compreensão do negócio:** Esta fase foca em compreender os objetivos do projeto e os requisitos do negócio, o que é crucial para definir o problema de forma adequada e estabelecer os objetivos da modelagem.
- **Entendimento dos dados:** Esta fase abrange a coleta inicial de dados e ações para entender as propriedades e a qualidade dos dados disponíveis, preparando o caminho para análises subsequentes.
- **Preparação dos dados:** Consiste na construção do conjunto de dados final a ser usado nas análises posteriores, tratando de questões como limpeza de dados, tratamento de valores ausentes e transformações necessárias.
- **Modelagem:** Nesta etapa, técnicas estatísticas e de aprendizado de máquina são aplicadas para construir modelos preditivos que possam efetivamente responder às questões do negócio.
- **Avaliação:** Os modelos são rigorosamente avaliados para garantir que atendam aos critérios de sucesso estabelecidos no entendimento do negócio. Ajustes são feitos conforme necessário para refinar o modelo.
- **Implantação:** Embora a implementação do modelo em um ambiente operacional não seja abordada neste estudo, propõe-se que futuras investigações possam explorar esta fase, onde as predições podem ser utilizadas para tomar decisões práticas ou para continuar refinando o sistema com novos dados coletados.

Com a metodologia definida, as próximas subseções detalharão cada fase, explorando as técnicas específicas utilizadas e como cada uma contribui para o alcance dos objetivos deste projeto.

3.1 Compreensão do negócio

A fase inicial deste estudo foca em compreender os requisitos do negócio em relação ao processo de secagem de lodo biológico. O principal objetivo é desenvolver um modelo preditivo que estime com precisão a consistência do lodo biológico na saída do secador de correia. Este entendimento surge da necessidade de melhorar a previsibilidade e o controle do processo, visando manter a consistência do produto final. As discussões com a equipe operacional e de gestão forneceram *insights* sobre as métricas de desempenho críticas e as limitações do processo atual. Estas conversas ajudaram a definir o escopo do projeto, delineando um caminho claro para o desenvolvimento de soluções analíticas que possam informar a tomada de decisões e suportar melhorias operacionais futuras.

3.2 Entendimento dos dados

Para investigar a eficiência do processo de secagem de lodo biológico utilizando uma RNA do tipo MLP, foi essencial estabelecer uma base de dados robusta. Inicialmente, foram realizadas reuniões com especialistas das áreas operacionais para definir as variáveis de processo que influenciam o desempenho do secador.

Os dados necessários para o estudo foram coletados por instrumentos que medem continuamente diversos parâmetros do processo. Esses dados são enviados de forma cabeada ao Sistema digital de controle distribuído (SDCD), responsável pelo controle operacional da planta de tratamento de lodo biológico. O SDCD é integrado a um subsistema *OLE for process control* (OPC), que permite a comunicação padronizada entre dispositivos de diferentes fabricantes e sistemas de controle. Um exemplo dessa integração é o *Plant information management system* (PIMS), que atua como um historiador de dados, arquivando as informações processadas. Para a extração dos dados relevantes para este estudo, foi utilizada uma extensão do *software Exaquantum* junto ao *Excel*. Esta ferramenta permitiu selecionar especificamente os dados de interesse, configurando o período de tempo e o intervalo entre amostras desejados, e posteriormente exportar essas informações para uma planilha no formato `.xls`.

O conjunto de dados final continha 259.201 registros, de 01 de janeiro a 30 de junho de 2023, com uma amostragem a cada minuto, proporcionando uma granularidade adequada para analisar a dinâmica do processo. Foram consideradas 66 variáveis contínuas, incluindo a taxa de alimentação real do secador, umidade relativa do ar de entrada e saída do secador, temperaturas do ar quente para cada segmento, consistência de entrada e saída (variável alvo), e as velocidades das esteiras.

Essas variáveis foram determinadas como cruciais para o desenvolvimento do modelo preditivo porque cada uma delas influencia diretamente diferentes aspectos do processo de secagem. Por exemplo, a taxa de alimentação real do secador impacta a capacidade de processamento e a homogeneidade do lodo tratado; a umidade relativa do ar de entrada e saída é essencial para avaliar a eficiência da secagem; as temperaturas do ar quente são críticas para a transferência de calor necessária para a evaporação da umidade; a consistência de entrada e saída do lodo reflete a eficiência do processo; e as velocidades das esteiras afetam o tempo de residência do lodo no secador, influenciando a eficácia da secagem.

Durante o pré-processamento, foi realizada uma transformação na variável alvo, `AI59_PV`, que corresponde à umidade relativa na saída do secador. Esta variável foi convertida para teor de sólidos totais (%TS) empregando a seguinte relação, conforme ilustrado na Equação (3.1):

$$\%TS = 100 - \%rF \tag{3.1}$$

Além disso, a variável 'Intervalo' foi convertida para *datetime* e definida como índice do *DataFrame*, facilitando a análise temporal dos dados.

O pré-processamento também incluiu uma verificação para identificar dados faltantes. Comandos em *Python*, executados no *Google Colab*, permitiram listar os nomes das colunas, quantificar a amostra de dados e verificar os tipos de dados, assegurando que todas as informações estivessem corretas e completas.

Além das verificações de integridade dos dados, bibliotecas de programação *Python* especializadas foram empregadas para otimizar a manipulação e análise dos dados coletados. A biblioteca *Pandas* provou ser essencial, facilitando a leitura e manipulação de dados tabulares, como os exportados do *Excel*. Funcionalidades para ajuste de datas e reorganização de dados foram particularmente úteis. Ademais, a biblioteca *NumPy* foi crucial para suportar operações numéricas avançadas, sendo extremamente útil em transformações de dados e cálculos complexos.

A escolha do *Google Colab* e da linguagem *Python* para todas as etapas de coleta, pré-processamento e análise dos dados foi motivada pela familiaridade e competência desenvolvida durante o curso de ciência de dados. Essas ferramentas proporcionaram um ambiente flexível e robusto, facilitando a manipulação e análise dos dados em todas as fases da pesquisa.

A tabela a seguir apresenta as variáveis selecionadas para o estudo, com suas respectivas descrições e unidades de medida, conforme foram identificadas como cruciais para o desenvolvimento do modelo preditivo:

Tabela 1 – Descrição detalhada das variáveis utilizadas no modelo.

Variáveis	**Descrição**	**Unidade de Medida**
AI56A_PV	Umidade relativa do ar fresco	%rF
AI57A_PV	Umidade relativa do ar de exaustão	%rF
AI59_PV	Consistência de saída do secador	%rF
AI60_PV	Consistência de entrada do secador	%TS
AI64_PV	Consistência de saída desaguamento	%TS
TC56_PV	Temperatura do ar fresco	°C
TI57_PV	Temperatura do ar de exaustão	°C
TI58A_PV	Temperatura água de entrada secador calorimetro	°C
TI58B_PV	Temperatura água de saída secador calorimetro	°C
TC60_PV	Temperatura do ar segmento 01 inferior	°C
TC61_PV	Temperatura do ar segmento 01 superior	°C
TC62_PV	Temperatura do ar segmento 02 inferior	°C
TC63_PV	Temperatura do ar segmento 02 superior	°C
TC64_PV	Temperatura do ar segmento 03 inferior	°C

Continua na próxima página

Tabela 1 – *Continuação da página anterior*

Variáveis	Descrição	Unidade de Medida
TC65_PV	Temperatura do ar segmento 03 superior	°C
TC66_PV	Temperatura do ar segmento 04 inferior	°C
TC67_PV	Temperatura do ar segmento 04 superior	°C
TC68_PV	Temperatura do ar segmento 05 inferior	°C
TC69_PV	Temperatura do ar segmento 05 superior	°C
TC70_PV	Temperatura do ar segmento 06 inferior	°C
TC71_PV	Temperatura do ar segmento 06 superior	°C
TC72_PV	Temperatura do ar segmento 07 inferior	°C
TC73_PV	Temperatura do ar segmento 07 superior	°C
TC74_PV	Temperatura do ar segmento 08 inferior	°C
TC75_PV	Temperatura do ar segmento 08 superior	°C
TC76_PV	Temperatura do ar segmento 09 inferior	°C
TC77_PV	Temperatura do ar segmento 09 superior	°C
TC78_PV	Temperatura do ar segmento 10 inferior	°C
TC79_PV	Temperatura do ar segmento 10 superior	°C
TC80_PV	Temperatura do ar segmento 11 inferior	°C
TC81_PV	Temperatura do ar segmento 11 superior	°C
TC82_PV	Temperatura do ar segmento 12 inferior	°C
TC83_PV	Temperatura do ar segmento 12 superior	°C
TC84_PV	Temperatura do ar segmento 13 inferior	°C
TC85_PV	Temperatura do ar segmento 13 superior	°C
TC86_PV	Temperatura do ar segmento 14 inferior	°C
TC87_PV	Temperatura do ar segmento 14 superior	°C
TC88_PV	Temperatura do ar segmento 15 inferior	°C
TC89_PV	Temperatura do ar segmento 15 superior	°C
PC16_PV	Pressão gás GNCD saída secador	mmH2O
PC28A_PV	Pressão ar fresco	mmH2O
PC61_PV	Pressão do ar segmento 1	mmH2O
PC62_PV	Pressão do ar segmento 2	mmH2O
PC63_PV	Pressão do ar segmento 3	mmH2O
PC64_PV	Pressão do ar segmento 4	mmH2O
PC65_PV	Pressão do ar segmento 5	mmH2O
PC66_PV	Pressão do ar segmento 6	mmH2O
PC67_PV	Pressão do ar segmento 7	mmH2O
PC68_PV	Pressão do ar segmento 8	mmH2O
PC69_PV	Pressão do ar segmento 9	mmH2O

Continua na próxima página

Tabela 1 – *Continuação da página anterior*

Variáveis	Descrição	Unidade de Medida
PC70_PV	Pressão do ar segmento 10	mmH2O
PC71_PV	Pressão do ar segmento 11	mmH2O
PC72_PV	Pressão do ar segmento 12	mmH2O
PC73_PV	Pressão do ar segmento 13	mmH2O
PC74_PV	Pressão do ar segmento 14	mmH2O
PC28B_PV	Pressão do ar segmento 15	mmH2O
FI58_PV	Vazão calorimetro para secador	m^3/h
FC60_PV	Vazão de água desmi para secador	m^3/h
SC30_PV	Velocidade rolo *crusher* 263SC01M5	%
SC46_PV	Velocidade esteira superior SC01M1	%
SC47_PV	Velocidade esteira inferior SC01M2	%
SC90_PV	Velocidade Faca peletizador EE07M2	RPM
SC91_PV	Velocidade Peletizador EE07M1	RPM
LI09_PV	Nível câmara de transferência	%
LI20_PV	Nível câmara de transferência	%
21SF004_TAL	Taxa de alimentação real de lodo	kg/h

3.3 Preparação dos dados

A Análise exploratória de dados (AED) é essencial no processo de modelagem preditiva, fornecendo *insights* profundos sobre distribuições, tendências e relações entre as variáveis. Utilizaram-se diversas ferramentas estatísticas e gráficas para realizar uma investigação abrangente do conjunto de dados, com o suporte de bibliotecas *Python* como Pandas, *NumPy*, *SciPy*, *Seaborn* e *Matplotlib*.

Inicialmente, a análise descritiva básica foi realizada utilizando o método *describe()* do *Pandas* para obter um resumo estatístico das variáveis. Esta etapa é crucial para identificar padrões ou anomalias nos dados. Histogramas com curvas de *Kernel density estimation* (KDE) foram gerados para cada variável numérica usando o *Seaborn* e o *Matplotlib*, oferecendo uma visualização clara das distribuições.

Para avaliar a normalidade das distribuições, aplicou-se o teste de *Shapiro-Wilk*, utilizando a biblioteca *SciPy*. Este teste ajudou a verificar se as variáveis seguem uma distribuição normal. Gráficos de caixa foram criados para visualizar a dispersão e possíveis *outliers* em diferentes grupos de variáveis. A matriz de correlação foi explorada através de um *heatmap* com o *Seaborn*, facilitando a identificação de possíveis relações lineares entre as variáveis.

A filtragem dos dados foi realizada estabelecendo limites específicos para cada variável, utilizando a biblioteca *Pandas* para excluir valores que poderiam distorcer as análises subsequentes. Para capturar as dinâmicas temporais das variáveis e entender melhor suas tendências e variações, foram calculadas estatísticas rolantes. As janelas de análise foram ajustadas conforme necessário para garantir relevância e precisão dos dados analisados. Especificamente, a média, o mínimo, o máximo, o desvio padrão e a contagem foram computados em uma janela de 150 períodos, com um mínimo de 60 períodos por janela para assegurar a robustez estatística. Essas estatísticas foram posteriormente deslocadas para trás, alinhando-as com os eventos futuros a serem previstos.

A correlação entre as variáveis de entrada e a variável alvo também foi analisada. Gráficos de dispersão e análises temporais adicionais foram realizados para cada variável em relação à variável alvo, utilizando decomposições de séries temporais e testes de estacionariedade com *Statsmodels*, proporcionando uma análise mais complexa e detalhada.

Na etapa de preparação dos dados para análise subsequente, efetuou-se a distinção das variáveis independentes e dependentes. Posteriormente, realizou-se a padronização da escala das variáveis utilizando o método *StandardScaler* da biblioteca *Scikit-learn*. Este método ajusta a distribuição dos dados para que apresentem média igual a zero e desvio padrão igual a um. A Equação (3.2) apresenta a fórmula utilizada para a padronização dos dados:

$$z = \frac{(x - \mu)}{\sigma} \tag{3.2}$$

onde x representa o valor original do dado, μ é a média dos dados e σ é o desvio padrão. Este procedimento é imprescindível para algoritmos de aprendizado de máquina que são sensíveis à escala e distribuição dos dados. A análise de *outliers* realizada após o processo de padronização corroborou a eficácia do método, assegurando a preparação adequada do conjunto de dados para o treinamento dos modelos preditivos.

Essas etapas de AED, suportadas por ferramentas analíticas robustas, forneceram uma base sólida para o desenvolvimento subsequente de modelos preditivos, garantindo que as decisões de modelagem fossem baseadas em um entendimento completo dos dados e suas complexidades.

3.4 Modelagem

3.4.1 Definição da arquitetura do modelo

Nesta subseção, apresenta-se a configuração e implementação do modelo de aprendizado de máquina do tipo MLP, empregado para analisar a eficiência do processo de secagem de lodo biológico. Inicialmente, os dados foram preparados dividindo-os em conjuntos de treino e teste, utilizando a função *train_split* da biblioteca *sklearn*. Optou-se por alocar 25% dos dados para o conjunto de teste, a fim de avaliar o desempenho do modelo em dados não vistos durante o treinamento, uma prática padrão para testar a generalização de modelos de aprendizado de máquina.

A estrutura do MLP foi definida com o uso da biblioteca *Keras* e consistiu em várias camadas densas. A primeira camada densa foi projetada para receber dados com dimensão igual ao número de variáveis explicativas do estudo. As camadas ocultas e a quantidade de neurônios em cada camada, bem como as funções de ativação, foram otimizadas utilizando o *Keras Tuner* com a técnica *Hyperband*. Este método envolveu a seleção da quantidade de neurônios em cada camada oculta entre 16 e 256, e a escolha das funções de ativação entre *relu*, *sigmoid* e *tanh*.

A técnica *Hyperband* permitiu realizar a busca de maneira eficiente, treinando diversos modelos com diferentes combinações de hiperparâmetros e selecionando as melhores configurações com base na métrica *val_mae*. O *EarlyStopping* foi utilizado para prevenir o sobreajuste, interrompendo o treinamento se não houvesse melhoria na função de perda de validação após cinco épocas consecutivas.

A camada de saída do modelo possui um único neurônio com função de ativação *linear*, adequada para tarefas de regressão onde o objetivo é prever um valor contínuo, refletindo a consistência do lodo no processo de secagem. O modelo foi compilado com o otimizador *adam*, selecionado por sua eficiência em convergir rapidamente durante o treinamento, e a função de perda *Mean squared error* (MSE), com o MAE como métrica de avaliação, proporcionando uma medida clara e compreensível do desempenho do modelo.

Adicionalmente, empregou-se a validação cruzada *K-Fold*, com 5 divisões, para garantir uma avaliação rigorosa do modelo. Esta técnica envolve dividir o conjunto de treino em 5 partes, treinando o modelo em 4 delas e validando em uma parte restante, repetindo o processo para cada uma das partes. Essa abordagem ajuda a identificar a estabilidade e a robustez do modelo. O controle de parada antecipada (*early stopping)* foi configurado para interromper o treinamento se não houvesse melhoria na função de perda de validação após cinco épocas consecutivas, prevenindo o sobreajuste.

A utilização das bibliotecas *Keras* e *TensorFlow*, junto com *sklearn*, permitiu a implementação eficiente de todas as etapas da modelagem, desde a divisão dos dados até a avaliação final do modelo. A escolha dessas ferramentas e parâmetros foi fundamentada na busca por um equilíbrio entre eficácia computacional e precisão preditiva, visando produzir um modelo confiável e adequado ao contexto do estudo em questão.

3.4.2 Treinamento e validação do modelo

Na etapa de treinamento e validação, o modelo de MLP foi configurado para otimizar seu desempenho e assegurar sua capacidade de generalização. A validação do modelo envolve não apenas treiná-lo com os dados disponíveis, mas também testar sua eficácia em um conjunto de dados separado, o que é crucial para evitar o sobreajuste e garantir que o modelo funcione bem em condições novas e desconhecidas.

Inicialmente, implementou-se a técnica de parada antecipada durante o treinamento para prevenir que o modelo se ajustasse excessivamente aos dados de treino. Este mecanismo de controle interrompe o treinamento caso a perda de validação não apresente melhoria após um determinado número de épocas, especificamente cinco neste estudo. A utilização dessa técnica ajuda a manter a qualidade do modelo ao ajustar-se aos dados sem perder a capacidade de generalizar para novos dados.

O treinamento foi conduzido ao longo de 50 épocas, um número definido para permitir que o modelo aprendesse adequadamente dos dados sem correr o risco de memorização. Durante cada época, o modelo foi avaliado usando um conjunto de dados de teste. Este processo não só ajuda a monitorar e ajustar a performance do modelo ao longo do tempo, mas também oferece uma visão clara de como o modelo pode ser esperado para se comportar em aplicações práticas.

Para avaliar a eficácia do modelo, o MAE foi calculado tanto para os dados de treinamento quanto para os de teste em cada época. Este indicador é essencial para medir a precisão das previsões do modelo, com valores menores indicando uma maior precisão. A comparação do MAE entre os conjuntos de treino e teste oferece uma medida robusta da capacidade do modelo de generalizar além dos dados utilizados durante o treinamento.

A visualização do desempenho do modelo ao longo das épocas foi realizada utilizando a biblioteca *Matplotlib*, que permitiu a geração de gráficos para comparar o MAE do treinamento e da validação. Essa visualização é fundamental para ajustes finos no processo de modelagem, garantindo que o modelo desenvolvido seja tanto robusto quanto eficiente.

A construção e treinamento do modelo foram realizados utilizando bibliotecas especializadas como *Keras* e *TensorFlow*, que oferecem ferramentas avançadas para a modelagem de redes neurais profundas. Estas bibliotecas proporcionam uma interface amigável e flexível para definir e otimizar uma variedade de arquiteturas de aprendizado de máquina, tornando-as escolhas ideais para implementar e ajustar o modelo MLP descrito.

Essa abordagem metodológica para o treinamento e validação assegura que o modelo MLP desenvolvido seja capaz de oferecer previsões confiáveis e precisas, essenciais para a aplicação prática na análise da eficiência do processo de secagem de lodo biológico. A escolha das ferramentas e técnicas foi guiada por práticas recomendadas na ciência de dados, visando a obtenção de um modelo bem ajustado que se comporta de maneira ótima em condições variadas.

3.5 Avaliação do modelo

Nesta subseção, descreve-se a metodologia adotada para avaliar o desempenho do modelo de MLP desenvolvido para analisar o processo de secagem de lodo biológico. Avaliar adequadamente o modelo é crucial para verificar sua precisão e eficácia em prever novos dados, proporcionando *insights* importantes sobre sua aplicabilidade prática.

Inicialmente, o desempenho do modelo foi avaliado no conjunto de teste, que não foi utilizado durante o treinamento para assegurar uma avaliação imparcial. Duas métricas principais foram calculadas: a perda (*Loss*) e o MAE. Estas métricas fornecem uma indicação clara do erro de previsão do modelo, sendo o MAE uma medida direta da magnitude dos erros em termos absolutos.

Após a avaliação inicial, foram realizadas predições usando o conjunto de teste. Essas predições foram usadas para calcular o *Root mean squared error* (RMSE) e o coeficiente de determinação (R^2), que são indicadores essenciais da qualidade das previsões. A formulação matemática do RMSE é apresentada a seguir:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3.3}$$

onde y_i são os valores reais e $\hat{y}_i$ são os valores preditos pelo modelo, e N é o número total de observações. O RMSE fornece uma medida do erro médio quadrático das previsões. O R^2 é calculado como:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3.4}$$

onde $\bar{y}$ é a média dos valores reais. O R^2 indica quão bem as variáveis independentes explicam a variabilidade da variável dependente.

Para uma análise mais detalhada do desempenho, gráficos visuais dos resultados das predições foram criados. Primeiramente, um gráfico de linha foi utilizado para comparar os valores preditos com os valores reais ao longo do tempo, permitindo visualizar a precisão do modelo em diferentes pontos do conjunto de teste. Gráficos de dispersão também foram elaborados para ilustrar a relação entre os valores preditos e os reais, com uma linha de identidade indicando a perfeição na previsão. Além disso, um gráfico dos resíduos versus os valores preditos foi utilizado para investigar a existência de padrões nos erros de previsão, e um histograma dos resíduos ajudou a analisar a distribuição dos erros.

Adicionalmente, um gráfico *Quantil-Quantil* (Q-Q) foi utilizado para avaliar se os resíduos da previsão seguem uma distribuição normal, o que é uma suposição comum em muitos testes estatísticos que avaliam modelos de regressão.

As ferramentas e bibliotecas utilizadas para essas análises incluíram *Keras*, para o treinamento e avaliação do modelo; *Matplotlib* e *Seaborn*, para a visualização dos dados; *Pandas*, empregado na manipulação e preparação dos dados para análise; e *SciPy*, utilizada para análises estatísticas adicionais, como o gráfico Q-Q.

Este conjunto de métodos e ferramentas proporcionou uma avaliação abrangente do modelo, destacando sua eficácia e áreas para futuras melhorias. A combinação de métricas quantitativas e visualizações qualitativas permite uma compreensão profunda da performance do modelo no contexto real da aplicação para a qual foi desenvolvido.

A avaliação do modelo MLP envolveu a utilização de um conjunto de dados de teste separado, não visto pelo modelo durante o treinamento, para medir a precisão das previsões. Métricas de desempenho como o MSE e a RMSE foram calculadas usando *Scikit-learn*, proporcionando uma avaliação quantitativa da capacidade do modelo de prever a consistência de saída do lodo. A análise dos resultados permitiu identificar áreas de melhoria e ajustar a arquitetura do modelo conforme necessário.

3.5.1 Importância das características

A interpretabilidade do modelo é fundamental para entender como as características influenciam as previsões no modelo MLP. Para realizar essa análise, utilizou-se a biblioteca SHAP, que se baseia na teoria dos jogos cooperativos para determinar a importância de cada característica através de sua contribuição para a capacidade preditiva do modelo.

Inicialmente, procedeu-se com a instalação da biblioteca SHAP e a importação das funções necessárias. Em seguida, um objeto '*Explainer*' foi criado utilizando o modelo MLP já treinado, juntamente com uma amostra representativa dos dados de treinamento, selecionada para refletir a distribuição geral do conjunto de dados. Esse objeto é responsável por calcular os valores SHAP, que quantificam a contribuição de cada característica para as previsões do modelo.

Os valores SHAP foram então computados para uma amostra selecionada do conjunto de teste, permitindo uma análise detalhada da influência de cada característica. Para facilitar a interpretação dessas influências, foram elaborados dois tipos de gráficos:

- **Gráfico de barras de importância média das características**: Este gráfico apresenta a importância relativa de cada característica, ordenando as barras de acordo com o impacto médio sobre as previsões do modelo. Ele é útil para identificar rapidamente os principais fatores que influenciam as previsões.
- **Gráfico SHAP *Force***: Um gráfico foi elaborado para as primeiras 150 amostras do conjunto de teste, mostrando como cada característica afeta as previsões individuais. Este gráfico representa graficamente a força da influência de cada característica, onde cada linha colorida indica uma característica diferente, com o comprimento e a direção da linha representando a magnitude e a direção (positiva ou negativa) do impacto da característica. O efeito total é visualizado como a soma dessas influências, ajudando a entender as interações complexas entre as características.

A metodologia SHAP aumenta a transparência das previsões do modelo e reforça a confiança nos resultados, fornecendo uma base sólida para futuras iterações do modelo. A análise detalhada das influências das características orienta ajustes no modelo, permitindo decisões informadas baseadas em evidências claras sobre quais características são mais informativas e quais podem ser descartadas como menos relevantes para a tarefa preditiva.

3.5.2 Ajuste de hiperparâmetros

O ajuste de hiperparâmetros é uma etapa importante no desenvolvimento de modelos de aprendizado de máquina, destinada a otimizar os parâmetros do modelo que não são diretamente aprendidos durante o treinamento mas que influenciam significativamente o desempenho do modelo. Para o modelo de MLP utilizado neste estudo, o ajuste foi realizado utilizando a biblioteca *Keras Tuner*, uma ferramenta avançada para a automação e otimização desses parâmetros.

Após instalar a *Keras Tuner*, definiu-se uma função de construção do modelo que permitia a exploração de uma gama variada de configurações de hiperparâmetros. No modelo sequencial, ajustou-se a quantidade de neurônios em cada uma das três camadas, variando de 16 a 256 com incrementos de 16 neurônios. As funções de ativação para cada camada também foram parametrizadas, com opções entre '*relu*', '*sigmoid*', e '*tanh*', sendo '*relu*' a escolha padrão para a primeira camada. Este arranjo flexível permitiu a experimentação com diferentes arquiteturas de rede de maneira sistemática.

Utilizou-se o método *Hyperband* da *Keras Tuner* para a otimização dos hiperparâmetros. *Hyperband* é uma abordagem de otimização baseada em competição que treina uma grande quantidade de modelos com diferentes hiperparâmetros por um número reduzido de épocas e seleciona apenas os modelos mais promissores para treinamentos mais extensos. Este método visa encontrar o melhor conjunto de hiperparâmetros de forma eficiente, minimizando o MAE como critério principal. O processo explorou as configurações durante o treinamento do modelo com os dados, utilizando até 50 épocas para cada configuração.

Ao concluir o processo de busca, o melhor modelo identificado foi extraído e submetido a um treinamento adicional para validar e melhorar ainda mais seu desempenho. Os resultados foram visualizados através de gráficos, comparando o MAE do treinamento e da validação ao longo das épocas, o que facilitou a análise da evolução do erro e ajudou a confirmar a eficácia do modelo selecionado.

Este método de ajuste de hiperparâmetros não só aprimorou a precisão do modelo, mas também forneceu *insights* valiosos sobre a influência de diferentes configurações na performance do modelo, assegurando que a escolha final dos parâmetros fosse bem fundamentada e alinhada com os objetivos do estudo.

3.6 Considerações éticas

A presente pesquisa, não envolve diretamente dados humanos, mas manipula informações provenientes de processos industriais, cujo acesso é restrito e regulado pela política de confidencialidade da empresa parceira. A fim de preservar a integridade e a confidencialidade dos dados industriais, todas as informações coletadas e utilizadas neste estudo foram tratadas com o máximo cuidado para garantir a anonimização e a segurança.

Os dados utilizados na análise foram exportados do sistema PIMS da empresa e consistem em registros operacionais que não estão disponíveis ao público. Para evitar a identificação direta ou indireta da empresa, nomes e identificadores específicos de equipamentos e processos foram substituídos por códigos genéricos, sem comprometer a validade científica das análises realizadas. Este procedimento de anonimização garante a proteção da propriedade intelectual da empresa e respeita as diretrizes de confidencialidade acordadas.

Um acordo formal de colaboração entre a universidade e a empresa estabelece claramente as diretrizes para uso, análise e divulgação dos dados. Este acordo inclui disposições sobre a propriedade intelectual, direitos autorais e limitações na publicação dos resultados do estudo, assegurando benefícios mútuos e a promoção de melhorias operacionais e ambientais.

Em suma, a seção de considerações éticas reflete o compromisso com a condução responsável da pesquisa, respeitando os princípios éticos que governam o uso de dados restritos e garantindo que os direitos e interesses de todas as partes envolvidas sejam devidamente protegidos e preservados. O processo de revisão e aprovação pela empresa antes da publicação reforça o compromisso com a ética, a transparência e a responsabilidade.

3.7 Limitações do estudo

A pesquisa apresenta algumas limitações, como a disponibilidade de dados históricos da fábrica de celulose, a complexidade dos modelos preditivos e a generalização dos resultados para outras fábricas de celulose.

3.8 Considerações finais

O presente capítulo apresentou a metodologia que será utilizada para o desenvolvimento e aplicação de modelos preditivos no processo de secagem de lodo biológico em uma fábrica de celulose. Os resultados da pesquisa poderão contribuir para a otimização da operação do secador de lodo, com potencial para gerar economia de recursos e redução do impacto ambiental.

Apresentar e discutir as hipóteses de forma detalhada. Explicar como cada hipótese se relaciona com os objetivos do estudo e como planejo testá-las. Devo detalhar o raciocínio lógico que leva às hipóteses, apoiado pela revisão da literatura e pelos objetivos da pesquisa.

4 Resultados

Este capítulo apresenta os resultados alcançados através da aplicação de técnicas avançadas de modelagem preditiva no processo de secagem de lodo biológico em uma fábrica de celulose Kraft. Utilizando a metodologia CRISP-DM, modelos de aprendizado de máquina, especificamente redes neurais do tipo MLP, foram desenvolvidos para prever a consistência do lodo na saída do secador. Este enfoque responde às necessidades críticas de eficiência e sustentabilidade, destacadas na revisão da literatura, e reafirma o papel essencial da ciência de dados em promover operações industriais mais eficazes e responsáveis ambientalmente.

A relevância desta pesquisa é enfatizada no contexto da indústria de celulose e papel, um setor econômico e tecnológico crucial no Brasil. A otimização do processo de secagem do lodo não apenas atende às exigências regulatórias e ambientais, mas também oferece vantagens econômicas significativas, destacando a importância de soluções sustentáveis. Por meio deste estudo, são propostas novas perspectivas e abordagens metodológicas baseadas em dados para aprimorar a gestão de resíduos industriais, contribuindo tanto para a literatura existente quanto para as práticas operacionais.

Os resultados obtidos são discutidos em relação às hipóteses formuladas, com uma análise crítica que vincula os achados aos estudos mencionados na revisão da literatura, como o trabalho de (BRÖNDUM, 2022), que também exploram a modelagem preditiva em contextos de tratamento de resíduos. Esta análise não apenas testa a validade das hipóteses propostas, mas também explora implicações práticas e teóricas, sublinhando o impacto da modelagem preditiva na eficiência e sustentabilidade dos processos industriais.

4.1 Apresentação dos dados

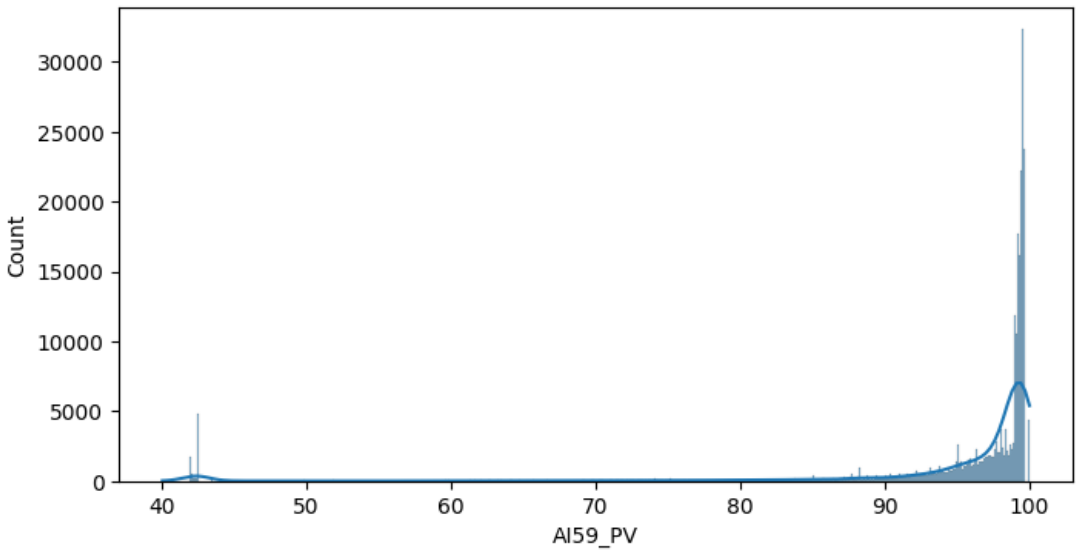
A análise inicial dos dados confirmou a robustez e integridade do conjunto utilizado para investigar o processo de secagem de lodo biológico. Composto por 259.201 registros e 66 variáveis de processo, o conjunto de dados não apresentou valores faltantes, destacando a alta qualidade da coleta de dados e a confiabilidade da base para análises subsequentes.

4.2 Preparação dos dados

A etapa inicial da AED revelou *insights* valiosos sobre a natureza dos dados coletados. A função `describe()` foi aplicada, fornecendo um resumo estatístico das variáveis, incluindo contagem, média, desvio padrão, mínimo, máximo e quartis para cada variável. As médias variaram significativamente entre as variáveis, indicando variações nos níveis operacionais e nas condições do processo. Desvios padrão altos em certas variáveis sugerem variabilidade substancial no processo, enquanto outras variáveis apresentaram desvios mais baixos, indicando maior consistência. Os valores mínimos e máximos refletiram os limites operacionais e as faixas de operação normal.

A investigação das distribuições variáveis esclareceu as condições operacionais habituais, facilitando a definição de limites para o aprimoramento dos dados. Concentrando-se na variável de saída, a consistência do lodo (AI59_PV), observou-se uma inclinação para leituras elevadas, reflexo dos controles que previnem a operação com níveis sub-ótimos de umidade. A variação acentuada na taxa de alimentação real do secador (21SF004_TAL) sugere oscilações no processo que poderiam ser atribuídas a múltiplos fatores, inclusive à consistência inadequada do lodo. Essa compreensão fornece um substrato de dados preparados para o desenvolvimento de modelos preditivos que visam intensificar a precisão e a eficiência operacional.

Figura 7 – Distribuição da variável alvo

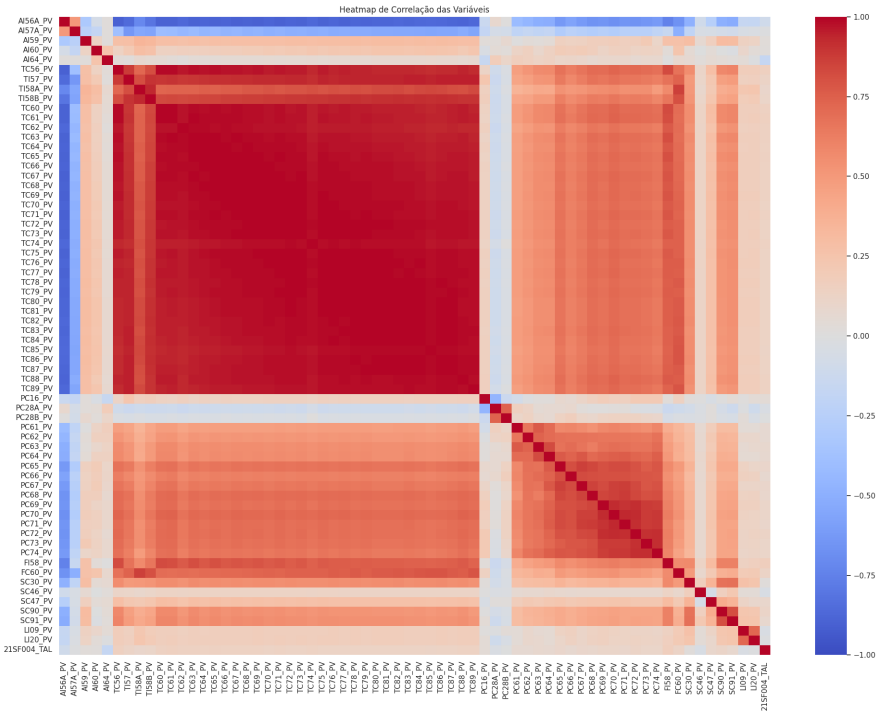


Fonte: elaborado pelo autor.

Os *boxplots* analisados ofereceram um panorama sobre a variabilidade e tendências centrais das variáveis críticas no processo. A técnica permitiu identificar os limites operacionais, destacar possíveis *outliers* e avaliar a consistência dos parâmetros. Com a variável alvo, a consistência de saída do lodo, apresentando uma tendência a valores mais altos, o que é alinhado com as expectativas operacionais e com os mecanismos de controle. Já a variabilidade notada na taxa de alimentação do secador sugere desafios na manutenção da estabilidade do processo.

A análise do *heatmap* foi fundamental para aprimorar o conjunto de dados a ser utilizado na modelagem preditiva. Identificou-se uma forte correlação entre as trinta variáveis de temperatura do secador (`TC60_PV` a `TC89_PV`), o que sugeriu uma condensação desses dados em uma única variável representativa da média, contribuindo para a redução de dimensões do modelo. Além disso, a alta correlação da umidade relativa de entrada (`AI56A_PV`) com as temperaturas indicou a possibilidade de sua exclusão, devido à sua natureza de indicador e não de controle, otimizando a relevância e a qualidade das variáveis preditoras.

Figura 8 – Heatmap de correlação das variáveis



Fonte: elaborado pelo autor.

A aplicação criteriosa dos filtros operacionais, orientada pelas análises de distribuição e correlação das variáveis, culminou em um aprimoramento da base de dados. As temperaturas dos segmentos do secador (`TC60_PV` a `TC89_PV`) foram consolidadas em uma única métrica média (`TCXX_MED`), racionalizando o conjunto de dados ao minimizar a redundância sem comprometer a integridade das informações. Esta condensação trouxe uma redução de dimensões significativa, essencial para um processamento e modelagem eficazes.

A eliminação de variáveis com alta quantidade de dados nulos ou inconsistências operacionais confirmadas, como `AI56A_PV` e `AI64_PV`, reforçou a robustez do conjunto. Mesmo com a remoção dessas variáveis, outras, consideradas críticas como a variável alvo (`AI59_PV`), foram preservadas apesar da descontinuidade observada, ressaltando a importância da precisão na preparação de dados. Essas decisões metodológicas resultaram em um *dataset* depurado, alinhado com o objetivo de construir um modelo preditivo eficiente, focado nas variáveis de maior relevância e controle operacional.

O tratamento estatístico avançado aplicado aos dados, incluindo a implementação de estatísticas rolantes, emergiu como uma abordagem eficaz na preparação do conjunto para a modelagem preditiva. A utilização de janelas de tempo adequadas, considerando o tempo de residência do secador, permitiu a captura de tendências e padrões temporais significativos, refletidos nos dados transformados.

As correlações reveladas pelos *heatmaps* após a aplicação das estatísticas rolantes reafirmaram a pertinência das variáveis de média e desvio padrão, ao passo que a exclusão de métricas mínimas, máximas e de contagem, as quais apresentaram *heatmaps* vazios, demonstrou a necessidade de focar em estatísticas que efetivamente contribuíssem para o poder preditivo do modelo. A exclusão criteriosa baseou-se na relevância operacional das variáveis e em análises preliminares, resultando em um *dataset* otimizado.

O refinamento adicional dos dados, com a eliminação de variáveis com alta presença de valores nulos e aquelas com multicolinearidade identificada, fortaleceu o conjunto de dados. As estatísticas rolantes introduziram uma nova dimensão de análise temporal, potencializando a detecção de dinâmicas processuais que variáveis estáticas não capturariam. Esta etapa foi crucial para assegurar que as variáveis remanescentes fossem representativas e robustas para a aplicação no modelo preditivo.

Em síntese, as etapas de tratamento e análise de dados, incluindo filtragens e estatísticas rolantes, cumpriram com o propósito de refinar o *dataset*, fundamentais para a construção de um modelo preditivo preciso e confiável. Os *heatmaps* subsequentes, focados nas variáveis de média e desvio padrão, confirmaram a eficácia dessas estatísticas na captura das relações entre as variáveis do processo, pavimentando o caminho para análises preditivas mais profundas e a implementação de controles operacionais automatizados no futuro.

Na etapa conclusiva de preparação de dados para a modelagem preditiva, procedeu-se a uma análise de correlação e à visualização gráfica subsequente ao tratamento estatístico avançado. As variáveis 'AI60_PV_original', 'AI57A_PV_mean' e 'AI57A_PV_std' emergiram como as mais significativamente correlacionadas com a variável dependente 'AI59_PV_shifted'. Gráficos de dispersão salientaram as relações intervariáveis, e a decomposição sazonal trouxe à tona padrões e ritmos latentes, evidenciando a consistência e capacidade de previsão inerente à série temporal.

O teste de *Dickey-Fuller* assegurou a estacionariedade da variável 'AI59_PV_shifted', corroborando a homogeneidade estatística e solidificando a base para um modelo preditivo robusto. A padronização dos dados por meio da técnica '*StandardScaler*' equiparou as variáveis explicativas, um passo imprescindível para as análises analíticas avançadas e detecção de *outliers*.

Observou-se a presença de possíveis *outliers*, conforme indicado pelos *boxplots*. Todavia, devido às intervenções antecedentes nos dados, optou-se por conservar tais pontos extremos, visando a manutenção da plenitude informacional. Esta resolução sublinha a necessidade de uma abordagem holística e prudente, especialmente considerando as particularidades dos dados industriais.

Como resultado, o processo de refinamento dos dados resultou na criação de uma base sólida para a próxima fase de modelagem preditiva. A precisão das análises anteriores garante que o conjunto de dados entregue ao modelo seja representativo e bem organizado, melhorando a precisão e a eficácia na previsão. Esse procedimento metodológico é fundamental para o sucesso na implementação de controles operacionais automatizados e para a evolução constante dos procedimentos industriais.

4.3 Modelagem preditiva

No desenvolvimento de modelos preditivos via redes neurais artificiais (RNA), especificamente o Perceptron Multicamadas (MLP), a definição cuidadosa da arquitetura da rede é essencial para otimizar o desempenho do modelo. A parametrização envolve estabelecer o número de camadas ocultas, o número de neurônios em cada camada, e as funções de ativação apropriadas, além de escolher o algoritmo de otimização apropriado.

Para este estudo, utilizou-se a biblioteca *Keras* do *Python* para construir o modelo MLP. A estrutura da rede consiste em:

- **Primeira camada oculta (Camada de entrada) e a segunda camada Oculta:** Compostas por 224 neurônios, que recebem os dados padronizados (*X_std*). Ambas com funções de ativação '*relu*' foi escolhida por sua eficiência e capacidade de solucionar o problema dos gradientes desaparecidos em redes profundas.
- **Terceira Camada Oculta:** Com a função de ativação '*sigmoid*'. A terceira camada oculta possui 176 neurônios. A função '*sigmoid*' foi escolhida por sua propriedade de normalizar a saída dos neurônios entre 0 e 1, o que pode ser particularmente útil para problemas de classificação.
- **Camada de Saída:** Consiste em um único neurônio com uma função de ativação '*linear*', adequada para problemas de regressão, indicando que o modelo está configurado para prever valores contínuos.

O modelo foi compilado utilizando o otimizador '*adam*', que é eficaz para grandes volumes de dados e variáveis devido à sua capacidade de ajustar a taxa de aprendizagem. A função de perda MSE e a métrica MAE foram selecionadas para quantificar o desempenho do modelo durante o treinamento.

Para determinar os melhores hiperparâmetros, pode-se utilizar o *Keras Tuner*, uma biblioteca de ajuste de hiperparâmetros para o *Keras*. O *Keras Tuner* realiza várias corridas de treinamento para avaliar o desempenho de diferentes combinações de hiperparâmetros, como a taxa de aprendizagem, a quantidade de neurônios e o número de camadas ocultas. Com base em uma função objetivo, como a minimização da função de perda, o *Keras Tuner* seleciona a melhor configuração de hiperparâmetros para o modelo.

A tabela a seguir ilustra os hiperparâmetros selecionados e seus respectivos valores:

Tabela 2 – Hiperparâmetros selecionados para o modelo de rede neural.

Hiperparâmetro	**Valor Selecionado**
Neurônios Camada 1	224
Neurônios Camada 2	224
Neurônios Camada 3	176
Função Ativação 1	relu
Função Ativação 2	relu
Função Ativação 3	sigmoid
Função Ativação Saída	linear
Otimizador	adam
Função de Perda	mse
Métrica	mae

Os achados na construção do modelo indicam que o MLP foi bem estruturado com uma complexidade suficiente para capturar a dinâmica dos dados sem ser excessivamente complexo, o que poderia levar a um sobreajuste. O uso de funções de ativação distintas nas camadas ocultas e na camada de saída foi intencional para refinar a capacidade do modelo de aprender padrões não-lineares nos dados. A seleção dos hiperparâmetros pelo *Keras Tuner* permitiu uma personalização fina do modelo, melhorando sua precisão e eficiência na predição dos resultados.

A definição cuidadosa dos parâmetros da Rede MLP é essencial para o sucesso do modelo preditivo. Essa etapa será fundamental para os resultados que serão apresentados na próxima seção do trabalho, na qual a performance do modelo será avaliada e discutida em detalhes.

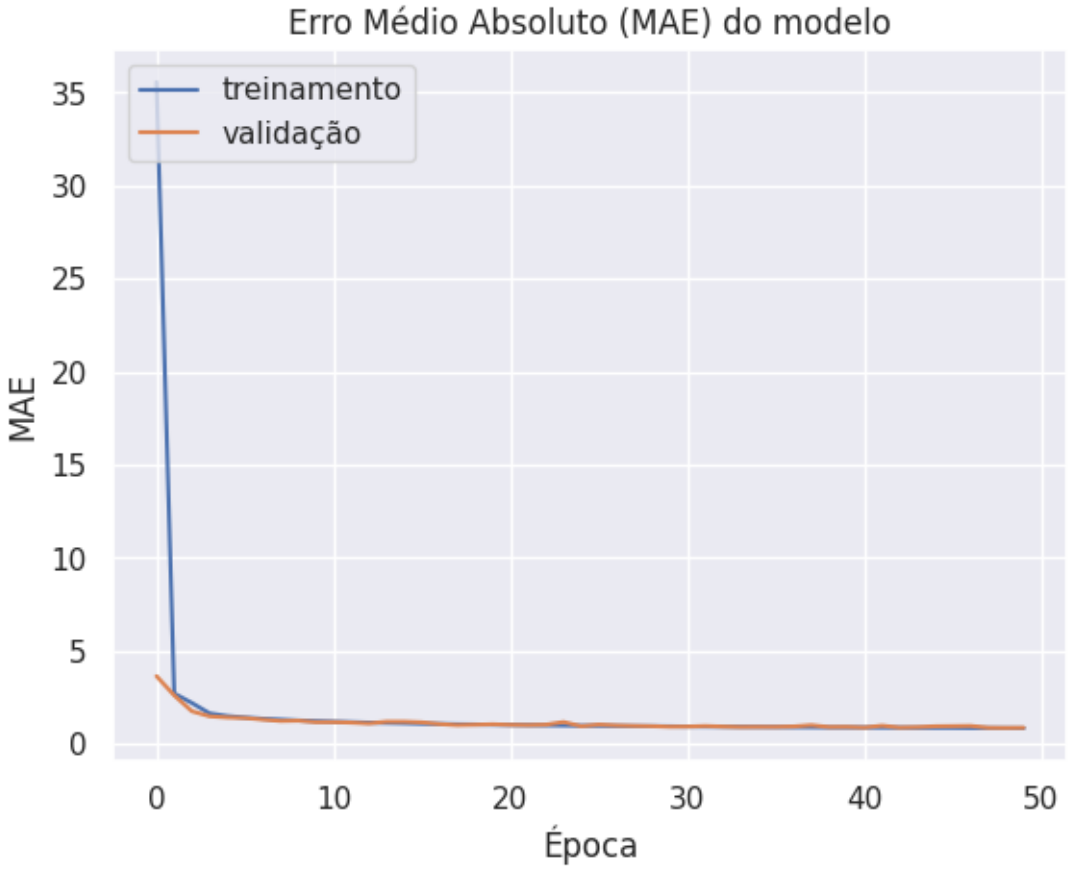
4.4 Avaliação e validação do modelo

A etapa de avaliação e validação do modelo desenvolvido foi crucial para assegurar sua eficácia prática. Empregando a técnica de validação cruzada *K-Fold* com $K = 5$, o modelo foi submetido a um teste que resultou em um MAE de 0,84%. Este valor indica que as previsões do modelo, em média, apresentam um desvio menor que uma unidade do valor real, conferindo uma precisão considerável para sua aplicação.

Um aspecto fundamental monitorado durante as etapas de treinamento e validação foi a prevenção do *overfitting*. Observou-se uma estabilização no MAE, evidenciando a capacidade do modelo em generalizar os dados ao invés de memorizar os padrões do conjunto de treinamento. A adoção da técnica de *Early Stopping* reforçou este comportamento, interrompendo o treinamento quando a melhoria do modelo cessou em um conjunto de validação independente.

A Figura 9 ilustra o comportamento do MAE ao longo das épocas para ambos os conjuntos de treinamento (azul) e validação (laranja). Nota-se que o MAE decresce rapidamente nas primeiras épocas e então se estabiliza, evidenciando a efetividade do processo de treinamento e a aplicação bem-sucedida de *Early Stopping* para impedir o sobreajuste do modelo.

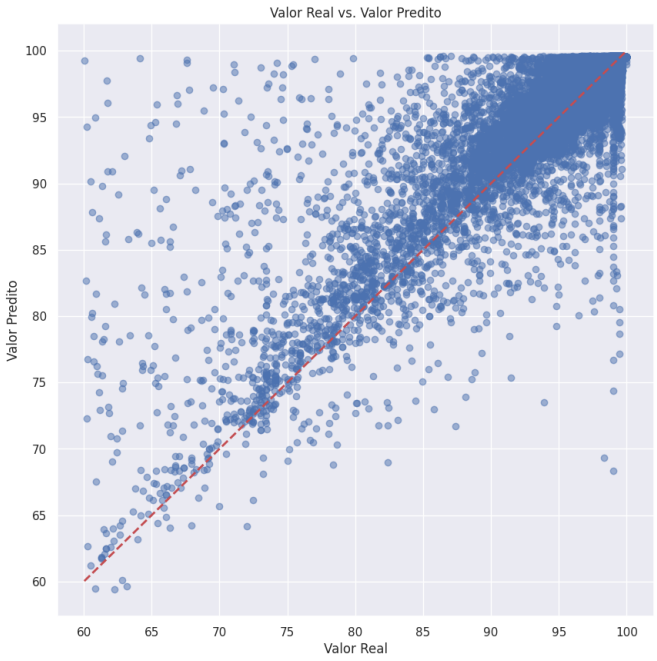
Figura 9 – Desempenho da RNA no treinamento e Validação.



Fonte: elaborado pelo autor.

A Figura 10 mostra a comparação entre os valores reais e os valores preditos pelo modelo. Os pontos próximos à linha diagonal vermelha indicam uma alta precisão nas previsões. A distribuição uniforme dos pontos ao redor desta linha sugere que o modelo mantém uma consistência na precisão das previsões através de diferentes faixas de valores, destacando a sua capacidade de generalização.

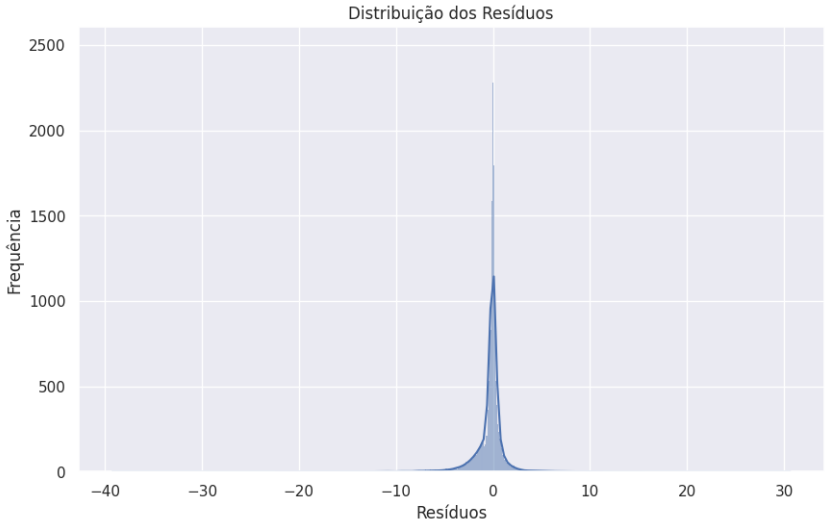
Figura 10 – Comparação do Valor Real e Predito



Fonte: elaborado pelo autor.

A Figura 11 apresenta a distribuição dos resíduos. A forma aproximadamente simétrica e centrada em torno de zero sugere que os erros do modelo são normalmente distribuídos, indicando que o modelo se ajusta bem aos dados sem apresentar erros sistemáticos significativos.

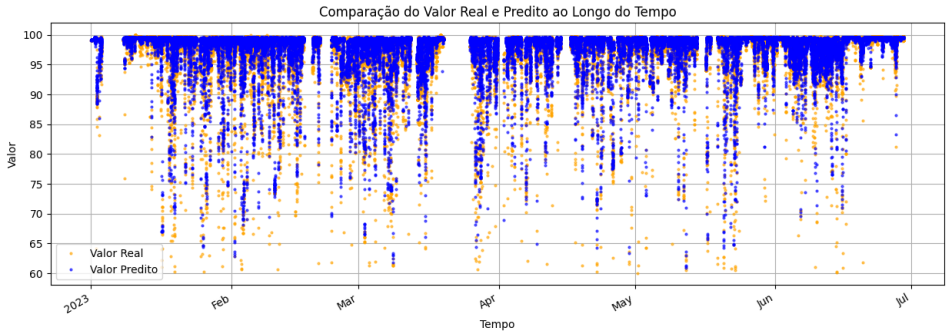
Figura 11 – Distribuição dos Resíduos



Fonte: elaborado pelo autor.

A habilidade do modelo em capturar a dinâmica dos dados ao longo do tempo é destacada pelo gráfico comparativo dos valores reais e preditos, mostrado na Figura 12. Apesar de algumas discrepâncias pontuais, uma forte correlação entre as previsões do modelo e os dados reais é evidenciada pelo coeficiente de determinação R^2 de 0,78, indicando que o modelo explica 78% da variância dos dados, o que reflete uma precisão substancial. Adicionalmente, o RMSE de 2,10, que corresponde a aproximadamente 46% do desvio padrão da variável-alvo de 4,59, revela que os erros de previsão, embora presentes, são moderados em relação à variabilidade total dos dados. Esses resultados sublinham que o modelo, apesar dos erros, mantém uma alta precisão relativa, sendo capaz de fazer previsões confiáveis com um nível de erro justificável dada a variabilidade natural dos dados.

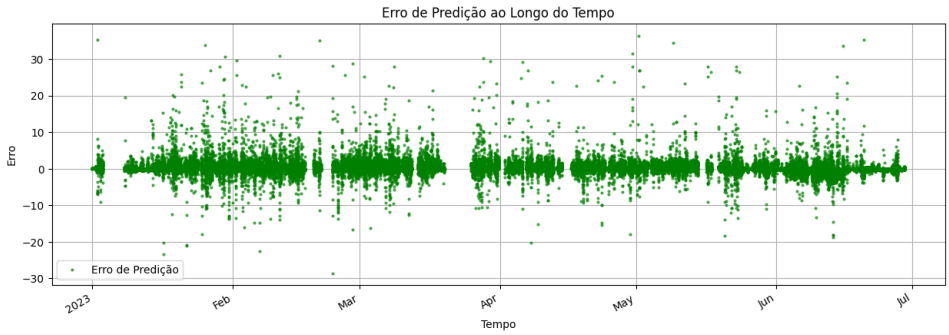
Figura 12 – Comparação do Valor Real e Predito ao Longo do Tempo



Fonte: elaborado pelo autor.

No segundo gráfico, mostrado na Figura 13, o erro de predição é plotado ao longo do mesmo intervalo temporal. O padrão de distribuição dos resíduos é fundamental para diagnosticar a consistência do modelo. A concentração de resíduos em torno de zero e a ausência de padrões estruturados sugerem que o modelo não é afetado por viés sistemático ou falhas na captura de tendências dos dados.

Figura 13 – Erro de Predição ao Longo do Tempo



Fonte: elaborado pelo autor.

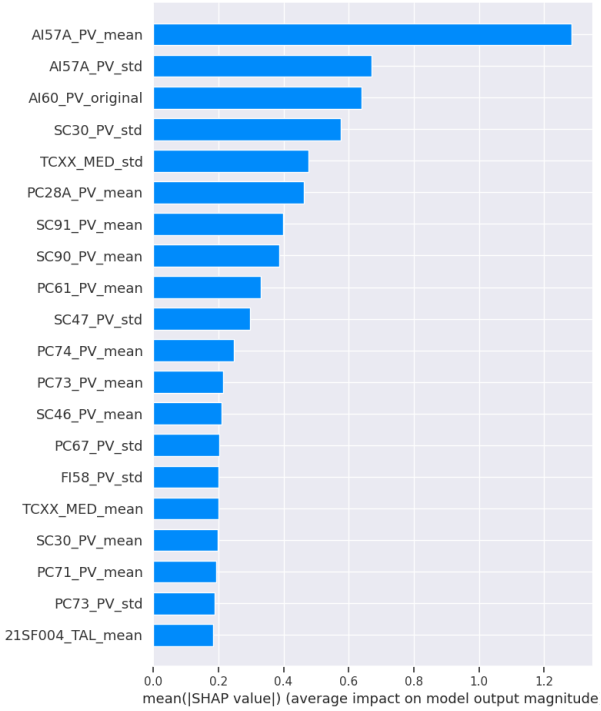
A avaliação gráfica reforça a competência do modelo de prever com precisão os valores ao longo do tempo e a análise dos resíduos confirma a adequabilidade do modelo para a tarefa de predição. Estes resultados realçam o potencial da aplicação do modelo para inferências futuras e apontam para a possibilidade de sua aplicação em um contexto prático.

4.5 Importância das características

A análise da importância das características foi realizada utilizando a metodologia SHAP, que oferece uma abordagem baseada em teoria dos jogos para quantificar a influência de cada variável nas previsões do modelo. As características *AI57A_PV_mean*, *AI57A_PV_std*, e *AI60_PV_original* foram identificadas como as mais influentes. Os gráficos SHAP, particularmente os *force plots*, forneceram *insights* sobre como essas variáveis afetam as saídas do modelo em diferentes condições operacionais.

A Figura 14 ilustra a importância relativa de cada característica. Este gráfico é fundamental para identificar quais variáveis contribuem mais para a variabilidade nas previsões e devem ser foco de análise detalhada para otimizações futuras do modelo ou para uma compreensão aprofundada dos mecanismos do processo modelado.

Figura 14 – Gráfico de barras mostrando a importância média das características.



Fonte: elaborado pelo autor.

Os gráficos *Force Plot* revelam diferenças significativas na forma como cada característica influencia as previsões do modelo.

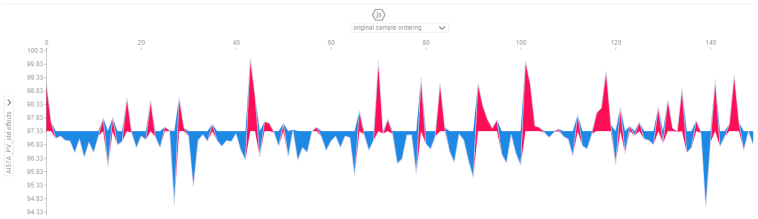
A Figura 15 mostra uma série de influências tanto positivas quanto negativas sobre a previsão do modelo. As contribuições positivas e negativas são relativamente balanceadas, com picos significativos que sugerem momentos onde a média dessa variável (*AI57A_PV_mean*)tem um impacto decisivo na previsão. Essa variável parece ter um papel dinâmico nas previsões, afetando-as de maneira substancial em pontos específicos.

Figura 15 – *Force plot* para AI57A_PV_mean mostrando seu impacto dinâmico nas previsões.

Fonte: elaborado pelo autor.

Em contraste com a figura anterior, a Figura para *AI57A_PV_std* apresenta picos mais acentuados em comparação com a média, indicando que a variabilidade (desvio padrão) de AI57A_PV tem uma influência mais marcante em certos pontos do modelo do que sua média. Isto sugere que o modelo é sensível não apenas aos valores centrais da AI57A_PV, mas também à sua dispersão. Os picos altos indicam que em momentos onde a variabilidade é maior, há um impacto considerável, seja positivo ou negativo, nas previsões do modelo.

Figura 16 – *Force plot* para AI57A_PV_std ilustrando a influência da variabilidade da característica nas previsões.



Fonte: elaborado pelo autor.

A Figura 17 apresenta uma relação complexa, com flutuações agudas e frequentes entre impactos positivos e negativos. Essa característica parece ter um comportamento mais volátil em termos de contribuição para as previsões do modelo, o que pode refletir uma interação complexa com outras variáveis ou uma resposta mais sensível a condições específicas do modelo.

Figura 17 – *Force plot* para AI60_PV_original destacando a complexidade de sua contribuição para o modelo.

Fonte: elaborado pelo autor.

Ao observar os gráficos *force plot* para as variáveis *AI57A_PV_mean* e *AI57A_PV_std*, nota-se uma variação significativa nos impactos que cada uma tem nas previsões do modelo. Especificamente, os picos mais acentuados no gráfico de *AI57A_PV_std* podem indicar que o modelo é extremamente sensível a mudanças na variabilidade desta variável. Uma possível explicação para essa sensibilidade pode ser que o modelo esteja captando não apenas tendências centrais, mas também variabilidades que indicam condições instáveis ou anômalas que são cruciais para a precisão das previsões em certos contextos operacionais. Este comportamento sugere uma análise adicional sobre como o modelo processa esses dados, o que poderia levar a *insights* sobre a natureza das dinâmicas modeladas e como elas reagem a variações mais sutis ou extremas nas entradas.

Estas análises revelam que, embora todas as três características sejam influentes, a natureza de sua influência varia significativamente. A média de AI57A_PV mostra o impacto mais dinâmico, a variabilidade (desvio padrão) de AI57A_PV tem um efeito mais moderado, e AI60_PV_original é notável por sua contribuição volátil e imprevisível. Compreender essas diferenças é crucial para otimizações futuras e ajustes no modelo, assim como para a implementação prática do mesmo em cenários aplicados.

4.6 Ajustes de hiperparâmetros

O ajuste de hiperparâmetros foi essencial para maximizar o desempenho do modelo. Utilizou-se o *Keras Tuner* com a técnica *HyperBand* para explorar sistematicamente várias configurações de hiperparâmetros, buscando a combinação ótima que minimizasse o erro de predição.

A Figura 18 mostra o processo de otimização dos hiperparâmetros, ilustrando como diferentes configurações afetaram o desempenho do modelo. A análise dos resultados permitiu selecionar a configuração que resultou em um MAE significativamente reduzido, destacando a eficácia do método *HyperBand* para ajuste rápido e eficiente.

Figura 18 – Resultados da otimização de hiperparâmetros mostrando o melhor modelo encontrado.

```
Trial 90 Complete [00h 08m 31s]
val_mae: 2.5863614082336426

Best val_mae So Far: 0.8493693470954895
Total elapsed time: 02h 47m 37s
```

```
# Resumo do melhor modelo encontrado
tuner.results_summary(1)
```

```
Results summary
Results in keras_tuner_dir/melhor_model
Showing 1 best trials
Objective(name="val_mae", direction="min")

Trial 0072 summary
Hyperparameters:
neurons_camada1: 224
neurons_camada2: 224
neurons_camada3: 176
ativacao_camada1: relu
ativacao_camada2: sigmoid
tuner/epochs: 50
tuner/initial_epoch: 17
tuner/bracket: 2
tuner/round: 2
tuner/trial_id: 0068
Score: 0.8493693470954895
```

Fonte: elaborado pelo autor.

Essas subseções consolidam a análise detalhada tanto da influência das características quanto da eficácia dos ajustes de hiperparâmetros, fornecendo uma base sólida para a compreensão do desempenho do modelo e indicando caminhos para melhorias futuras.

O processo de otimização de hiperparâmetros, realizado com o auxílio do *Keras Tuner* e a técnica *HyperBand*, foi decisivo para aprimorar a acurácia do modelo. A seleção criteriosa dos hiperparâmetros levou a uma melhora nos indicadores de performance, com a obtenção de um MAE inferior nos conjuntos de teste. Isso não apenas valida a metodologia adotada, mas também assegura a confiabilidade do modelo para aplicações futuras, evidenciando a sua capacidade de adaptação e aprimoramento contínuo.

Cada fase da avaliação e validação do modelo contribuiu para um entendimento mais abrangente da sua capacidade preditiva e das áreas que ainda podem ser melhoradas. A otimização constante desses processos é essencial para avançar em direção a um modelo ainda mais confiável e eficiente.

5 Considerações Finais

Este trabalho explorou métodos de aprendizado de máquina para prever comportamentos em processos industriais específicos. Utilizando técnicas avançadas, como a validação cruzada *K-Fold* e a técnica SHAP, foi possível melhorar a confiabilidade e a interpretabilidade dos modelos preditivos desenvolvidos. Essas técnicas provaram ser eficazes, fornecendo modelos que podem melhorar significativamente a tomada de decisão e otimizar operações.

Os modelos preditivos demonstraram uma capacidade notável de fornecer previsões precisas, crucial para setores como a manufatura. A integração desses modelos em sistemas de análise preditiva pode ajudar na otimização da cadeia de produção, redução de desperdícios e melhoria na resposta a variações de mercado, elevando a competitividade das empresas. Essas implicações práticas sublinham o potencial dos modelos de aprendizado de máquina em promover operações mais eficientes e sustentáveis, reforçando a importância de tais tecnologias na indústria moderna.

No entanto, as limitações deste estudo devem ser reconhecidas. Os resultados obtidos podem não ser totalmente generalizáveis devido à especificidade dos conjuntos de dados utilizados e à complexidade dos modelos, que podem não ser aplicáveis em cenários que demandam soluções mais simples e interpretação direta. Além disso, a eficácia dos modelos está condicionada à escolha de hiperparâmetros ótimos, e a implementação prática requer avaliações adicionais, particularmente em termos de integração de sistemas e desempenho em ambiente operacional real.

Para futuras pesquisas, é recomendável a exploração de conjuntos de dados mais amplos e diversificados para melhorar a generalização e robustez dos modelos. Também seria benéfico investigar abordagens de modelagem que equilibrem precisão e interpretabilidade, o que é essencial para aplicações em que decisões rápidas e claras são cruciais. A realização de testes práticos em ambientes operacionais reais fornecerá *insights* sobre o desempenho e adaptabilidade dos modelos, ajudando a transformar o potencial teórico em benefícios tangíveis para as indústrias.

Em resumo, este estudo não apenas confirmou a eficácia das técnicas de aprendizado de máquina na previsão e otimização de processos industriais, mas também abriu caminho para futuras investigações no campo do aprendizado de máquina e controle preditivo. As contribuições deste trabalho estabelecem uma base sólida para a continuação da pesquisa, promovendo a inovação contínua e o aumento da eficiência operacional e sustentabilidade em diversas indústrias.

Referências

ALPAYDIN, E. **Introduction to Machine Learning**. 4. ed. Cambridge, MA, USA: MIT Press, 2020.

AMBIENTE, C. N. do M. **Resolução CONAMA nº 430 de 2010, que dispõe sobre condições e padrões de lançamento de efluentes**. 2010.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York, NY, USA: Springer, 2006.

BRASIL. **Política Nacional de Resíduos Sólidos**. 2010. <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/lei/l12305.htm>. Lei Nº 12.305, de 2 de agosto de 2010.

BRÖNDUM, E. **Modeling of the primary sludge thickening process at a wastewater treatment plant with the use of machine learning**. 2022.

CLIFTON, E.; THURAISINGHAM, R. A framework for business intelligence. **Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-34)**, p. 1–10, 2001.

COSTA, D. S. D.; CRISTELLI, F. P.; PATROCINIO, A. B. D. A new conception to biosludge treatment and destination in kraft pulp mills. In: . [s.n.], 2021. v. 82, n. 2, p. 82 – 91. Cited by: 0. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163549188&partnerID=40&md5=ecf6e61fab32c47171b8d33bc1ff7ded>.

EIKELBOOM, M.; LOPES, A.; SILVA, C.; RODRIGUES, F. A multi-criteria decision analysis of management alternatives for anaerobically digested kraft pulp mill sludge. **PLoS One**, Public Library of Science, 2018.

FAHIM, S.; NISAR, N.; AHMAD, Z.; ASGHAR, Z.; SAID, A. Managing paper and pulp industry by-product waste utilizing sludge as a bio-fertilizer. **Polish Journal of Environmental Studies**, Polish Journal of Environmental Studies, 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA, USA: MIT Press, 2016.

HINTON, G. E.; SEJNOWSKI, T. J. **Unsupervised Learning: Foundations of Neural Computation**. Cambridge, MA, USA: MIT Press, 1999.

Huber Technology. **Imagem de Tecnologia de Tratamento de Água**. 2023. <https://www.huber-technology.com/fileadmin/_processed_/7/f/csm_bild_titel_bt_03_e412dd6b00.jpg>. Acessado em 2 de maio de 2024.

HUILIñIR, C.; ALLEN, D.; TRAN, H. Drying characteristics of biosludge from pulp and paper mills. **University of Toronto Libraries**, University of Toronto, 2017.

IBÁ, I. B. D. **Relatório Anual 2023**. São Paulo: Indústria Brasileira de Árvores, 2023. 88 p. Acessado em: [12 abr. 2024].

KDnuggets. **Polls: Data Mining Methodology**. 2007. <https://www.kdnuggets.com/2007/polls/data-mining-methodology>. [Online; accessed 20-April-2024].

LIMEIRA, J. Environmental management of sludge in pulp and paper industry. **SpringerBriefs in Applied Sciences and Technology**, Springer, 2016.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>.

LUNDBERG, S. M.; LEE, S.-I. **Front Page Example (XGBoost)**. 2021. Accessed: 2024-06-26. Disponível em: <https://shap-lrjball.readthedocs.io/en/latest/example_notebooks/tree_explainer/Front%20page%20example%20(XGBoost).html?highlight=shap.plots.force#Front-page-example-(XGBoost)>.

METZNER, J. Advanced effluent treatment processes in pulp and paper industry. **Journal of Industrial Pollution Control**, Techno Press, v. 34, n. 2, p. 2152–2163, 2018.

MITCHELL, T. M. **Machine Learning**. New York, NY, USA: McGraw-Hill, 1997.

MURPHY, K. P. **Machine learning: a probabilistic perspective**. [S.l.]: MIT press, 2012.

NAVAEE-ARDEH, S.; BERTRAND, F.; STUART, P. R. Emerging biodrying technology for the drying of pulp and paper mixed sludges. **Drying Technology**, Taylor & Francis, v. 24, n. 7, p. 863–878, 2006.

PINHEIRO, A. C. M. **A indústria de papel e celulose no Brasil: um estudo de sua competitividade no cenário internacional**. Tese (phdthesis) — Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

PéREZ-ENCISO, M.; ZINGARETTI, L. M. A guide for using deep learning for complex trait genomic prediction. **Genes**, v. 10, n. 553, p. 1–19, 2019. Disponível em: <www.mdpi.com/journal/genes>.

RAMOS, J. L. C.; RODRIGUES, R. L.; SILVA, J. C. S.; OLIVEIRA, P. L. S. de. Crisp-edm: uma proposta de adaptação do modelo crisp-dm para mineração de dados educacionais. In: SBC. **Anais do XXXI Simpósio Brasileiro de Informática na Educação**. [S.l.], 2020. p. 1092–1101.

SANTOS, A.; VAZ, T.; LOPES, D.; CARDOSO, O. Beneficial use of lime mud from kraft pulp industry for drying and microbiological decontamination of sewage sludge. **Journal of Environmental Management**, Elsevier, 2021.

SEADI, T. A.; RUTZ, D. Bioenergy from anaerobic degradation of lipids in palm oil mill effluent. **Renewable Energy**, Elsevier, v. 129, p. 776–784, 2019.

SHOKRI, R.; SHMATIKOV, V. Privacy-preserving deep learning. In: ACM. **Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security**. [S.l.], 2015.

SILVA, C.; ROSA, E. Energy recovery from sludge and sustainable waste management. **Journal of Environmental Management**, Elsevier, v. 183, p. 777–784, 2016.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning: An Introduction**. 2. ed. Cambridge, MA, USA: MIT Press, 2018.

TOCZYŁOWSKA-MAMIńSKA, R. Limits and perspectives of pulp and paper industry wastewater treatment–a review. **Renewable and Sustainable Energy Reviews**, Elsevier, v. 78, p. 764–772, 2017. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1364032117306561>.

WIRTH, H.; HIPP, J. Crisp-dm: A process model for knowledge discovery. ACM Press, New York, p. 276–280, 2000.