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

Modelo preditivo de alvura final em estágio de branqueamento de polpa celulósica

Rhuan Duarte Carvalho

João Monlevade, MG
2024

Rhuan Duarte Carvalho

Modelo preditivo de alvura final em estágio de branqueamento de polpa celulósica

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. Alexandre Magno de Sousa

João Monlevade, MG

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

C331m Carvalho, Rhuan Duarte.
Modelo preditivo de alvura final em estágio de branqueamento de polpa celulósica. [manuscrito] / Rhuan Duarte Carvalho. - 2024.
77 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Alexandre Magno de Sousa.
Produção Científica (Especialização). Universidade Federal de Ouro Preto. Departamento de Engenharia de Produção.

1. Indústria de papel e celulose. 2. Otimização de processos. 3. Aprendizado de máquina. I. Sousa, Alexandre Magno de. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.85:676.02

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

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Rhuan Duarte Carvalho

Modelo Preditivo de Alvura Final em Estágio de Branqueamento de Polpa Celulósica

Monografia apresentada 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 5 de julho de 2024.

Membros da banca

Prof. Dr. Alexandre Magno de Sousa - Orientador (UFOP)
Prof. Dr. Luiz Carlos Bambirra Torres - (UFOP)
Me. Yoni Armando Minchola Robles - (Suzano)

Professor Alexandre Magno de Sousa, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 16/07/2024.

Documento assinado eletronicamente por **Alexandre Magno de Sousa**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 23/07/2024, às 15:38, 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 **0744990** e o código CRC **C7FE5130**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

À minha amada esposa, Sofia Elisa, minha fonte constante de inspiração e apoio incondicional. Sua presença ao meu lado tornou esta jornada acadêmica mais leve, pois sempre acreditou em mim e compartilhou cada passo desta jornada.

Agradecimentos

Quero expressar minha profunda gratidão a todos que tornaram possível a realização deste trabalho, pois cada contribuição foi essencial para o sucesso.

Ao Professor Alexandre Magno, meu orientador dedicado, agradeço pela paciência, sabedoria e orientação ao longo deste processo. Sua expertise foi fundamental para a qualidade deste trabalho.

Ao Yoni Robles, meu superior e gerente de processos, agradeço pela confiança depositada em mim e pelo constante apoio para me permitir concluir essa pós-graduação, que foi vital para o desenvolvimento desta pesquisa.

Ao Igor Favaro, amigo e pessoa estagiária, agradeço pela parceria e pelo esforço conjunto na realização deste trabalho. Sua colaboração foi valiosa e imprescindível o resultado final.

À Caroline Vasconcelos, cientista de dados da Digital da Suzano, agradeço pelo apoio contínuo, conselhos preciosos e pelo compartilhamento de conhecimentos, que foram fundamentais para o enriquecimento desta pesquisa.

À Cenibra, que possibilitou a realização desta pós-graduação, expresso meu profundo agradecimento pelo investimento em educação e desenvolvimento profissional.

À Suzano, agradeço pela oportunidade e pelo suporte que me permitiu conciliar o trabalho com a pós-graduação, contribuindo significativamente para a conclusão deste projeto.

A todos, meu sincero obrigado por fazerem parte deste percurso e por tornarem possível a concretização deste trabalho acadêmico.

E tudo que pensei
E tudo que eu falei
E tudo que me contaram
Era papel.

E tudo que descobri
Amei
Detestei
papel.
CARLOS DRUMMOND DE ANDRADE

Resumo

O objetivo deste trabalho é demonstrar a viabilidade de modelos preditivos de alvura em estágios de branqueamento de celulose e identificar as principais variáveis de entrada úteis para essa predição. O trabalho foi desenvolvido na empresa Suzano Papel e Celulose, unidade de Aracruz – ES, líder global na produção de celulose de eucalipto e uma das maiores fabricantes de papéis no mercado internacional. Esta pesquisa busca avaliar a eficácia de diferentes técnicas de aprendizado de máquina com foco na aplicação prática desses modelos em processos industriais para melhorar a precisão das previsões de alvura, a qual é um parâmetro crítico para a qualidade do produto final. Foram aplicados três modelos de aprendizado de máquina, a saber: Regressão Linear, XGBoost, e LightGBM. Os resultados mostram que os métodos do estado da arte como XGBoost e LightGBM alcaçaram um R^2 de 0.80 e são 25% melhores que a Regressão linear. Isso demonstra a superioridade de modelos baseados em árvores de decisão para capturar relações não lineares e complexas entre as variáveis. O modelo XGBoost foi utilizado para aplicação da previsão da alvura no PI System que coleta e gerencia dados da planta provenientes dos sistemas PLC e SDCD. Essa integração demonstra a viabilidade da aplicação de modelos de predição avançados em processos de produção em tempo real. Por fim, a aplicação do modelo no sistema contribui para a otimização do processo, redução de custos e, principalmente, na melhoria da qualidade do produto final.

Palavras-chaves: Produção de celulose. Cozimento Kraft. Predição de Alvura. Otimização Industrial. Machine Learning.

Abstract

The objective of this work is to demonstrate the feasibility of predictive models for brightness in cellulose bleaching stages and to identify the main input variables useful for this prediction. The work was developed at Suzano Papel e Celulose, Aracruz unit – ES, a global leader in eucalyptus pulp production and one of the largest paper manufacturers in the international market. This research aims to evaluate the effectiveness of different machine learning techniques with a focus on the practical application of these models in industrial processes to improve the accuracy of brightness predictions, which is a critical parameter for the quality of the final product. Three machine learning models were applied, namely: Linear Regression, XGBoost, and LightGBM. The results show that state-of-the-art methods like XGBoost and LightGBM achieved an R^2 of 0.80 and are 25% better than Linear Regression. This demonstrates the superiority of tree-based models in capturing nonlinear and complex relationships between variables. The XGBoost model was used for the application of brightness prediction in the PI System, which collects and manages plant data from PLC and DCS systems. This integration demonstrates the feasibility of applying advanced prediction models in real-time production processes. Finally, the application of the model in the system contributes to process optimization, cost reduction, and, most importantly, improving the quality of the final product.

Keywords: Pulp production. Kraft Cooking. Brightness Prediction. Industrial Optimization. Machine Learning.

Lista de ilustrações

Lista de tabelas

Lista de Abreviaturas

AM	Aprendizado de Máquina 1, 4, 14, 16
CDF	Cumulative Distribution Function 22, 23
CV	Coeficiente de Variação 22
DFT	Discrete Fourier Transform 16
ECF	Elemental Chlorine Free 5, 7, 17
GBDT	Gradient Boosting Decision Trees 12
HexA	Ácido Hexenurônico c, 2, 8, 20, 21, 23, 25, 26, 40
IMAVol	Incremento Médio Anual Volumétrico 16
KNN	k-Nearest Neighbor 15
LDA	Linear Discriminant Analysis 16
LPF	Licor Preto Fraco 6
MAE	Mean Absolute Error 4, 11, 14, 32, 33, 41, 42, 44, 45, 47, 48, 51–53, 57, 58
MSE	Mean Square Error 4, 11, 14, 32, 33, 41, 42, 44, 45, 47, 48, 51–53, 57, 58
PCA	Principal Component Analysis 16, 19, 29–32, 40, 54, 61
PEC	Pulsed Eddy Current 16
PI System	Plant Information System 4, 35, 54–57
PLC	Programmable Logic Controller 54
RF	Random Forestr 17
RMSE	Root Mean Square Error 15, 17

R^2	Coeficiente de determinação 4, 14, 15, 17, 32–34, 41–49, 51–54, 57, 58
SDCD	Sistema Digital de Controle Distribuído 54
SVM	Support Vector Machines 15, 16

Sumário

1 Introdução

A indústria de papel e celulose desempenha uma função importante na produção de materiais essenciais para diversos setores da sociedade, por isso, é um *player* relevante na economia global (PROPEQ, 2022). A constante busca por avanços tecnológicos dentro desse setor visa não apenas atender às demandas do mercado, mas também otimizar processos operacionais, reduzir custos e minimizar o impacto ambiental associado às atividades industriais (SÖDERHOLM; BERGQUIST; SÖDERHOLM, 2019). Diante disso, a polpação *kraft* destaca-se como o principal método de produção de celulose e o branqueamento, processo subsequente, uma etapa essencial na obtenção das especificações de alvura exigidas pelos clientes (SANTOS; HART, 2013).

O branqueamento da celulose, posterior ao cozimento, depuração e deslignificação, é composta geralmente de três a seis estágios e destaca-se como uma fase crítica no processo, o que demanda um controle preciso para garantir a qualidade desejada da celulose. De forma geral, cada estágio do branqueamento apresenta desafios ao controle, principalmente o primeiro estágio, pelo fato de absorver a maior parte da variação na entrada do processo (COLODETTE *et al.*, 2005). A alvura da polpa pode ser descrita como uma medida de quão branca está a polpa ao longo do processo de branqueamento e a necessidade de desenvolver estratégias inovadoras para melhorar o controle é evidenciada pela importância atribuída a esse parâmetro na indústria de papel e celulose.

Nesse contexto, surge a proposta deste trabalho científico de desenvolver um sistema de predição de alvura para um estágio de branqueamento, utilizando técnicas avançadas de Aprendizado de Máquina (AM). O objetivo é fornecer às fábricas de celulose uma ferramenta que permita antecipar a alvura da celulose com precisão, contribuindo para melhorias substanciais na eficiência operacional, na gestão ambiental e na qualidade do produto final.

A importância estratégica desse trabalho fundamenta-se na constante busca por inovações no setor, que tem a alvura da celulose, como parâmetro crítico, que influencia diretamente na satisfação do cliente e a competitividade das empresas (PAULA, 2022). A implementação de um sistema de predição permitirá um melhor aprimoramento no controle de dosagem de químicos em cada estágio do branqueamento (NOVEL, 2018).

1.1 Motivação e Justificativa

A motivação deste trabalho está na busca por estratégias inovadoras que possam melhorar o controle e a eficiência do processo de branqueamento de celulose. A alvura da celulose tem sido alvo de atenção especial devido à sua importância na satisfação do cliente e na competitividade das empresas do setor. A busca por métodos preditivos mais precisos, que operem de forma online, representa uma lacuna que este trabalho busca preencher.

A justificativa para este estudo é fundamentada na necessidade de desenvolver um modelo preditivo de alvura que possa proporcionar às fábricas de celulose uma ferramenta precisa e confiável para antecipar a alvura final a cada estágio de branqueamento. Tal modelo não apenas aprimoraria o controle da qualidade do produto, mas também contribuiria para a redução de custos operacionais e para o aumento da eficiência e uma gestão ambiental ainda mais responsável.

Os químicos usados no processo de branqueamento de celulose de fibra curta *kraft* costumam representar de 15% a 20% do custo caixa das fábricas. Esse valor é extremamente significativo e impactam na lucratividade e na competitividade das empresas. Em fábricas que utilizam o dióxido de cloro como agente químico branqueador, ele costuma representar sozinho metade do custo total de químicos de branqueamento.

Portanto, este trabalho se propõe a abordar essa problemática por meio do desenvolvimento e aplicação de um modelo preditivo de alvura para estágios de branqueamento, destacando sua relevância estratégica e as potenciais contribuições para o setor. Além disso, a implementação de técnicas avançadas de aprendizado de máquina neste contexto beneficia todo o setor, por meio de práticas industriais mais eficientes, sustentáveis e orientadas para a qualidade do produto.

1.2 Definição do Problema

A problemática central nesta pesquisa está intrinsecamente ligada à complexidade do controle da alvura da celulose no estágio de branqueamento, onde diversas variáveis impactam diretamente na qualidade do produto final. Entre essas variáveis, destacam-se a alvura de entrada da polpa, número *kappa*, teor de HexA, produção do branqueamento, tempo de reação, carga de químicos, pH, temperatura e teor de *shives*. A interação dinâmica dessas variáveis complexas representa uma lacuna que impacta diretamente a eficiência operacional, gestão ambiental e competitividade das empresas no setor.

A ausência de um modelo preditivo robusto dificulta a capacidade de antecipar com precisão os patamares de alvura final em cada estágio, dificultando a otimização do processo e resultando em produtos finais que não atendem plenamente às expectativas de qualidade exigida. A necessidade de uma ferramenta confiável que leve em consideração essas variáveis é evidenciada pela importância estratégica que a alvura desempenha na especificação de cada produto para cada cliente.

1.3 Objetivo Geral e Específicos

O objetivo geral deste trabalho é desenvolver e implementar um modelo preditivo de alvura para um estágio de branqueamento na indústria de celulose, utilizando técnicas avançadas de aprendizado de máquina. Para que este objetivo seja alcançado, os seguintes objetivos específicos são definidos:

- Realizar revisão da literatura sobre métodos de branqueamento, identificando desafios e abordagens tradicionais;
- Analisar e criticar as técnicas de controle de processo atuais na indústria, explorando oportunidades para inovações;
- Desenvolver e implementar um modelo preditivo de alvura usando técnicas de aprendizado de máquina para o estágio de branqueamento;
- Avaliar a eficácia do modelo preditivo, por meio dos valores preditos e observados durante a produção;
- Implementar nova estratégia para otimização do controle de dosagem de químicos no estágio de branqueamento.

1.4 Resultados e Contribuições

Este trabalho demonstrou a viabilidade de modelos preditivos de alvura em estágios de branqueamento de celulose, destacando a eficácia de técnicas de aprendizado de máquina para prever um parâmetro crítico na qualidade do produto final. Os resultados quantitativos para os tês modelos de aprendizado de máquina mostraram que os modelos baseados em árvores de decisão (*XGBoost* e *LightGBM*) superaram significativamente a Regressão Linear.

A aplicação da Regressão Linear para prever a alvura da polpa resultou em métricas de desempenho medianas, com um Mean Absolute Error (MAE) de 0.911, um Mean Square Error (MSE) de 1.324 e um Coeficiente de determinação (R^2) de 0.599. O MAE indica a magnitude média dos erros de predição, o MSE representa a média dos quadrados dos erros, refletindo a dispersão dos dados ao redor da linha de melhor ajuste, e o R^2 quantifica a proporção da variabilidade na variável dependente que é explicada pelo modelo. O modelo *XGBoost*, por outro lado, apresentou um desempenho superior em todas as métricas avaliadas em comparação com a Regressão Linear. Os resultados mostraram um MAE de 0.617, um MSE de 0.670 e um R^2 de 0.798. Com isso, O *XGBoost* demonstrou ser 32.27% melhor em MAE, 49.40% melhor em MSE e 24.94% melhor em R^2 em comparação com a Regressão Linear. Similarmente, o *LightGBM* também apresentou um desempenho superior em relação à Regressão Linear. Os resultados mostraram um MAE de 0.610, um MSE de 0.661 e um R^2 de 0.797. O *LightGBM* se destacou ao ser 33.04% mais preciso em MAE, 50.08% melhor em MSE e 24.84% superior em R^2, também em relação à Regressão Linear. Esses modelos capturaram com maior precisão as relações não lineares e complexas entre as variáveis independentes e a alvura, demonstrando que a relação não é perfeitamente linear, o que limita a eficácia da Regressão Linear.

As contribuições deste trabalho são interessantes tanto para a indústria de celulose quanto para a literatura acadêmica. Na prática industrial, a implementação do modelo *XGBoost* no Plant Information System (PI System) da unidade industrial mostrou a viabilidade de integrar modelos preditivos avançados em processos de produção em tempo real. Isso possibilita a otimização do processo, redução de custos variáveis e melhorias na qualidade do produto final. Para a literatura acadêmica, este estudo reforça a eficácia dos modelos baseados em árvores de decisão em cenários industriais complexos.

1.5 Estrutura e Organização da Monografia

O presente trabalho está estruturado conforme descrito a seguir. No Capítulo 2, são apresentados os principais conceitos do processo de branqueamento, incluindo uma revisão da literatura sobre métodos de branqueamento e a relevância da alvura na produção de celulose, bem como as abordagens tradicionais de controle de processo e a necessidade de inovações para enfrentar os desafios específicos de cada estágio de branqueamento (MATHUR *et al.*, 2018). O Capítulo 3 descreve a metodologia, incluindo a análise das variáveis da base de dados e os fundamentos teóricos e práticos das técnicas de AM a serem aplicadas, destacando sua aplicabilidade no contexto da predição de alvura. O Capítulo 4 apresenta a implementação dos modelos propostos, os resultados obtidos e as discussões sobre esses resultados. Finalmente, no Capítulo 5 são apresentadas as considerações finais e os possíveis trabalhos futuros.

2 Revisão da Literatura

Este capítulo apresenta uma revisão detalhada dos principais conceitos e processos envolvidos na fabricação de polpa celulósica, com foco especial no estágio de branqueamento. A seção aborda desde a importância econômica da indústria de papel e celulose no contexto brasileiro até os detalhes operacionais das etapas de cozimento *kraft*, branqueamento Elemental Chlorine Free (ECF) e métodos de aprendizado de máquina utilizados.

O capítulo inicia com uma visão geral do setor de papel e celulose no Brasil, destacando seu impacto econômico e estatísticas recentes. Em seguida, são detalhadas as etapas operacionais da fábrica de celulose, abrangendo a linha de fibras e a recuperação química, destacando apenas a importância de cada área do processo. Após isso, são discutidos o processo de cozimento contínuo *kraft* de fibras curtas, com ênfase nos componentes do licor branco de cozimento, o controle do número *Kappa* e a viscosidade da polpa. Em seguida, é abordado o branqueamento de celulose ECF, detalhando suas principais etapas e os agentes branqueadores utilizados. Ao final, é detalhado o principio de funcionamento dos modelos que serão utilizados: Regressão Linear, *XGBoost* e *LightGBM*.

2.1 Fundamentação Teórica

Segundo dados do último levantamento disponibilizado pela Indústria brasileira de árvores, em 2019, o setor de árvores plantadas no Brasil representou um faturamento de R$97,4 bilhões e as exportações somaram cerca de US$ 11,3 bilhões, o equivalente a 4,3% das exportações brasileiras. Neste anuário foi contabilizado que 36% do uso das árvores foi destinado ao setor de papel e celulose, representando um volume total superior a 19 milhões de toneladas de celulose produzida em todo país naquele ano (IBA, 2019).

Além do enorme peso que a área florestal representa na geração de valor, pela geração de mudas, plantio e colheita, a área industrial também possui grande impacto econômico, desde a demanda de consumíveis até a geração dos produtos e subprodutos do processo fabril (IBA, 2019).

Uma fábrica de celulose regular pode ser dividida em duas grandes áreas, a saber, a linha de fibras (áreas 1,2 e 3) e a recuperação química (áreas 4 e 5) conforme é apresentado na Figura 1. A linha de fibras é responsável pela extração da celulose e pode ser subdividida em: picagem de toras e armazenamento de cavacos (área 1); cozimento dos cavacos de madeira e branqueamento da polpa (área 2); secagem e enfardamento das folhas de celulose (área 3) (GOMIDE; GOMES, 2015). A recuperação química produz o licor branco de cozimento *kraft* a partir do próprio licor já reagido, denominado Licor Preto Fraco (LPF), e pode ser subdividida em: evaporação do LPF (área 4); queima do licor preto concentrado nas caldeiras; caustificação do licor verde; geração do licor branco (área 5) (REIS, 2021).

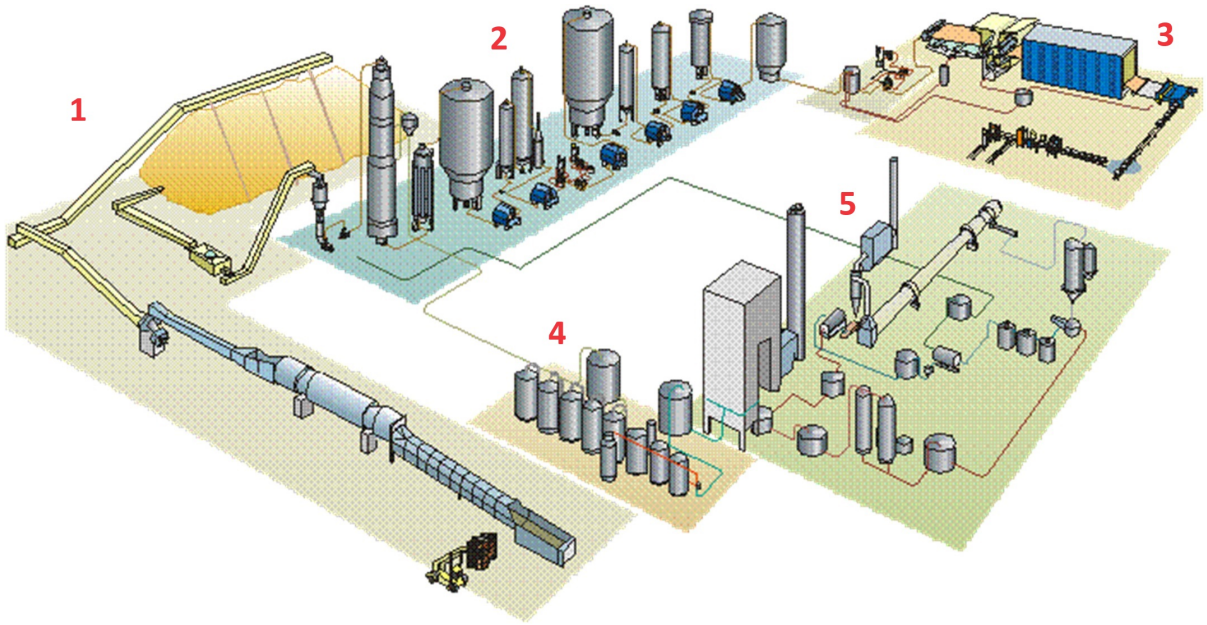


Figura 1 – Fluxograma de uma fábrica de celulose.

Fonte: Sandberg (2017), adaptada pelo autor.

O presente trabalho propõe a construção de um modelo de predição de alvura na etapa de branqueamento, utilizando métodos de aprendizagem de máquina aplicados a um problema de regressão. Esse processo é uma importante subdivisão da linha de fibras e é responsável por elevar a alvura da polpa descarregada do digestor até atendimento desse parâmetro aos clientes. Assim, nos tópicos a seguir, será apresentada uma revisão de literatura das áreas de cozimento *kraft* e branqueamento de celulose.

2.1.1 Cozimento contínuo *kraft* de fibras curtas

A etapa de cozimento dos cavacos para extração de celulose ocorre em equipamento denominado digestor, neste caso, um digestor contínuo de aproximadamente 60 metros de altura. Neste equipamento, os cavacos de madeira reagem com o licor branco de cozimento *kraft* e a lignina é solubilizada para liberação das fibras (FOELKEL, 2009).

O licor branco de cozimento *kraft* é composto por dois sais de sódio em meio aquoso, hidróxido de sódio (NaOH) e sulfeto de sódio (Na2S), na proporção que varia de 30% a 40% (p/p) desse último. O termo *kraft*, significa "forte" em alemão, pois este desenvolvimento representou um ganho de resistência das polpas em relação ao cozimento "soda". Além disso, a adição do sulfeto também trouxe melhora em outros parâmetros importantíssimos, tais como rendimento e branqueabilidade da polpa, dada sua maior seletividade em relação ao hidróxido de sódio (GOMIDE; GOMES, 2015).

O principal parâmetro de controle de um digestor é denominado *Kappa*. Segundo Correia, d'Angelo e Júnior (2019), o número *Kappa* é um método indireto para determinação da lignina residual e outros grupos da polpa. Rigorosamente, ele indica o número de equivalentes de oxidação com uma solução de permanganato de potássio ($KMnO_4$) sob condições ácidas precisamente especificadas.

Outro parâmetro que cabe mencionar nesse tópico da revisão é a viscosidade da polpa. A análise é feita com a dissolução da polpa de celulose em etilenodiamina cúprica (EDC) e a viscosidade dinâmica é determinada pelo tempo que esse fluido leva para percorrer através do capilar, sob a influência da gravidade, de um viscosímetro convencional. Essa medida está correlacionada ao grau de degradação da polpa nas cadeias de carboidratos (celulose e hemicelulose) e também pode ser interpretada como uma medida do rendimento da polpação (GOMIDE, 2002).

2.1.2 Branqueamento de celulose ECF

A etapa do processo que se seguem ao cozimento *kraft* é denominado branqueamento ECF, ele pode ser subdividido em: depuração da polpa marrom, deslignificação com oxigênio e branqueamento da celulose. Essas etapas estão relacionadas ao atingimento da maior parte dos parâmetros de qualidade da celulose para os clientes.

O processo de branqueamento ECF surgiu após um intenso movimento de órgãos ambientais e clientes que foi progressivamente proibindo e restringindo possibilidade de formação de cloro elementar no branqueamento. Ele consiste na utilização do dióxido de cloro (ClO_2) como agente branqueador em substituição ao gás cloro (Cl_2), essa mudança eliminou a geração de dioxinas e demais compostos clorados no efluente de branqueamento (VENTORIM *et al.*, 1999).

A depuração da polpa marrom consiste na operação unitária de separação mecânica para remoção dos incozidos de madeira, resilientes ao processo de cozimento, geralmente correlacionados aos nós dos galhos, as cascas das árvores e aos cavacos picados em sobretamanho (*oversizes*).

A deslignificação com oxigênio, também conhecida como pré-branqueamento, consiste na reação da polpa com o gás oxigênio em meio alcalino e contribui de forma significativa para redução do número kappa na entrada do branqueamento, a qual é uma tecnologia bem estabelecida e utilizada em larga escala (RABELO, 2006). Entretanto, essa redução é quase exclusivamente advinda da remoção de lignina residual da polpa, pois algumas estruturas ainda presentes na polpa não apresentam reatividade com o oxigênio, como os ácidos hexenurônicos (HexA) e os complexos lignina-carboidrato. Assim, a eficiência da deslignificação ficará sempre limitada à presença de tais estruturas (CAUX; DALVI; AMORIM, 2013).

A lignina é o principal componente cromóforo da polpa, ou seja, ela é capaz de conferir cor a celulose, ela é quantificada na polpa pelo número *Kappa*. No branqueamento, o número kappa permanece sendo um importante parâmetro, pois além da lignina, ele quantifica os ácidos hexenurônicos (HexA) e outros extrativos presentes na polpa. O HexA é formado durante o processo de cozimento a partir de grupos orgânicos ligados a Xilana e é removido da polpa por hidrólise em estágios ácidos do branqueamento. Ele possui um um grupo enol-éter (dupla ligação presente) conjugado a um grupo carboxílico, conforme demonstrado na Figura 2. Por isso, ele é considerado um grupo leuco-cromóforos, ou seja, sob determinadas condições ele não apresenta cor, mas sob outras condições ele pode passar a conferir cor a celulose (RABELO, 2006).

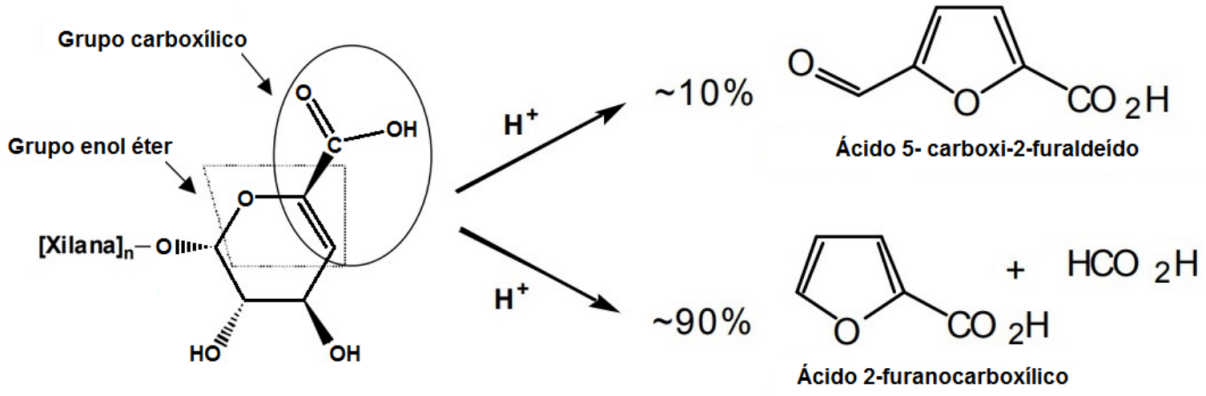


Figura 2 – Hidrólise ácida de HexA (grupo cromóforo da celulose).

Fonte: (CARVALHO *et al.*, 2019).

O branqueamento da celulose é feito pela reação da polpa com agentes oxidantes fortes, com alta capacidade de alvejamento ou enzimas altamente seletivas, o principal objetivo é realizar a remoção de lignina residual e ácidos hexenurônicos. Muitos agentes são utilizados atualmente como: dióxido de cloro, peróxido de hidrogênio, ozônio, oxigênio, hipoclorito de sódio, ácido peracético, enzima xilanase e outros (COLODETTE *et al.*, 2006).

Na Figura 3 é apresentado um modelo de planta de branqueamento de celulose com três estágios, semelhante ao branqueamento A, onde o trabalho está sendo desenvolvido. O primeiro estágio, denominado DHT, consiste na reação da polpa com o dióxido de cloro (ClO_2), em baixo pH aprox. 3,5, acidificado com ácido sulfúrico (H_2SO_4) e em alta temperatura aprox.90°C. O segundo estágio, denominado EP, consiste em uma extração alcalina com peróxido de hidrogênio (H_2O_2), alcalinizado com hidróxido de sódio (NaOH) e temperatura de aprox. 80°C. O terceiro e último estágio, denominado D1, consiste novamente na reação com dióxido de cloro (ClO_2), em pH moderado aprox. 5,5, também acidificado com ácido sulfúrico (H_2SO_4) e e temperatura de aprox. 80°C.

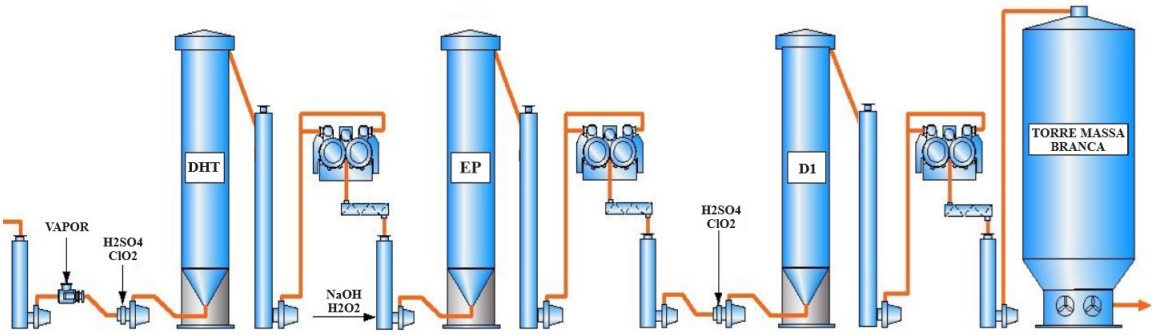


Figura 3 – Planta de branqueamento de celulose contendo três estágios DHT, EP e D1.

Fonte: (RAGNAR; BACKA, 2004), adaptada pelo autor.

Cada estágio de reação é seguido de uma etapa de lavagem para remoção dos produtos de oxidação. Os equipamentos de lavagem podem ser prensas, filtros atmosféricos, filtros pressurizados, difusores, prensas e outros, que embora variem a tecnologia e a eficiência, todos se prestam ao mesmo trabalho. Uma planta de branqueamento geralmente possui de 3 a 6 estágios, plantas mais antigas com equipamentos de lavagem menos eficientes operam com mais estágios e as mais novas com menos.

A alvura da polpa pode ser descrita como uma medida de quão branca e luminosa está a polpa ao longo do processo de branqueamento e é geralmente expressa em unidades de Brilho Ótico (*ISO brightness*). Na metodologia de medição utiliza-se um filtro azul e mede-se a refletância da luz no comprimento de onda de 457 nanômetros, dentro do espectro visível que vai de 380 nm (violeta) até 770 nm (vermelho). Dessa forma, a refletância da polpa é medida em comparação com um padrão branco ideal (PEDRO, 2017).

Os produtos podem conter especificações variadas, podendo incluir alvura, reversão de alvura máxima, sujidade, microkappa, pH, teor de cinzas e outras. Os produtos *High brightness* tem níveis de alvura superiores a 91,5% ISO, os *Extra Prime* superior a 88,5% ISO e os *Low brightness* superior a 85,0% ISO. O preço do produto aumenta nas maiores especificações de alvura e índice de sujidade máximo permitido também.

2.1.3 Regressão Linear

O modelo de regressão linear é um dos métodos mais simples e amplamente utilizados na análise de dados e aprendizado de máquina (SANTOS, 2018). Seu objetivo principal é modelar a relação entre uma variável dependente (ou alvo) e uma ou mais variáveis independentes (ou preditoras) utilizando uma linha reta (ou hiperplano no caso de múltiplas variáveis preditoras).

A equação fundamental da regressão linear simples (com uma única variável preditora) é (MACARRINGUE, 2022):

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{2.1}$$

em que:

- y é a variável dependente;
- x é a variável independente;
- β_0 é o intercepto do modelo, representando o valor esperado de y quando x é zero;
- β_1 é o coeficiente de regressão, indicando a mudança esperada em y para uma unidade de mudança em x;
- ε é o termo de erro, representando a diferença entre os valores observados e os valores previstos pelo modelo.

Na regressão linear múltipla, existem várias variáveis preditoras ($x_1, x_2, \ldots, x_n$) e a equação é expandida para:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon. \tag{2.2}$$

A regressão linear faz várias suposições sobre os dados, incluindo a linearidade, independência, a homoscedasticidade e normalidade dos erros (ROCHA *et al.*, 2015):

- Linearidade: a relação entre as variáveis dependente e independente é linear;
- Independência: as observações são independentes umas das outras;
- Homoscedasticidade: a variância dos erros é constante em todos os níveis da variável independente;
- Normalidade dos Erros: os erros (resíduos) seguem uma distribuição normal.

Cada coeficiente de regressão β_j (onde j varia de 1 a n) representa a mudança esperada na variável dependente para uma mudança de uma unidade na variável independente correspondente, mantendo todas as outras variáveis constantes.

A qualidade do ajuste do modelo é frequentemente avaliada usando métricas como o coeficiente de determinação (R^2), que indica a proporção da variabilidade total em y que é explicada pelo modelo. Outras métricas incluem o erro quadrático médio (MSE) e o erro absoluto médio (MAE).

A regressão linear, apesar de sua simplicidade e ampla aplicabilidade, possui tanto vantagens quanto limitações que devem ser consideradas ao utilizá-la. Entre suas principais vantagens e desvantagens estão:

- Vantagens: simplicidade, interpretabilidade, eficiência computacional e ampla aplicação;
- Limitações: sensibilidade a *outliers*, incapacidade de modelar relações não lineares e a suposição de que as variáveis independentes não são altamente colineares.

A regressão linear é uma técnica fundamental em estatística e aprendizado de máquina, conhecida por sua simplicidade e interpretabilidade (SANTOS, 2023). Ao aplicar a regressão linear, é crucial verificar as assunções do modelo e considerar possíveis limitações, especialmente em casos de relações não lineares ou presença de multicolinearidade entre variáveis preditoras.

2.1.4 *XGBoost*

O *XGBoost*, ou *Extreme Gradient Boosting*, é uma implementação eficiente e escalável do algoritmo de *Gradient Boosting* para árvores de decisão. Desenvolvido por Chen *et al.* (2015), é uma técnica de aprendizado de máquina supervisionado amplamente utilizada em problemas de regressão e classificação, com diversas aplicações práticas.

O princípio básico do *XGBoost* é combinar várias árvores de decisão fracas para formar um modelo preditivo robusto. Cada árvore é construída sequencialmente, com o objetivo de cada algoritmo corrigir os erros do anterior.

O *XGBoost* implementa um processo conhecido como *boosting*, que consiste em combinar vários algoritmos fracos para produzir um algoritmo forte. A ideia do *boosting* é utilizar os algoritmos fracos de forma sequencial, onde cada um busca minimizar o erro do algoritmo anterior. Durante o treinamento, o algoritmo atribui pesos aos exemplos de treinamento com base nos erros residuais, dando mais atenção aos exemplos mal classificados.

As principais características do *XGBoost* incluem:

- Regularização: o *XGBoost* suporta diferentes técnicas de regularização para evitar *overfitting*, incluindo a regularização L1 (Lasso) e L2 (*Ridge*);
- Otimização de Perdas: o algoritmo permite a otimização de uma variedade de funções de perda, como erro quadrático médio (MSE) para problemas de regressão e entropia cruzada para problemas de classificação;

- Tratamento de Dados Ausentes: o *XGBoost* possui um mecanismo embutido para lidar com valores ausentes nos dados, evitando a necessidade de pré-processamento adicional;
- Paralelismo: o algoritmo é altamente paralelizável, o que permite treinamento rápido em conjuntos de dados grandes;
- Suporte a GPU: o *XGBoost* pode ser executado em unidades de processamento gráfico (GPU), acelerando ainda mais o treinamento em grandes conjuntos de dados.

O *XGBoost* é conhecido por sua eficácia e desempenho superior em competições de ciência de dados e é amplamente utilizado na indústria e na academia devido à sua precisão e escalabilidade. Ele se tornou uma escolha popular para uma variedade de problemas de aprendizado de máquina devido à sua flexibilidade e capacidade de lidar com dados complexos.

2.1.5 *LightGBM*

O *LightGBM* é um algoritmo de aprendizado de máquina baseado em árvore de decisão que também utiliza o método de *gradient boosting*. Esta abordagem permite lidar com grandes conjuntos de dados de forma eficiente e com desempenho superior em relação a outros métodos de árvores de decisão. O *LightGBM* é um algoritmo de código aberto que utiliza árvores de decisão com reforço de gradiente, chamados de Gradient Boosting Decision Trees (GBDT).

O princípio de funcionamento do *LightGBM* é semelhante ao do *XGBoost*, onde várias árvores de decisão fracas são combinadas para formar um modelo preditivo robusto. No entanto, o *LightGBM* introduz algumas otimizações que o tornam mais rápido e eficiente. Ele acelera o treinamento, reduz o consumo de memória e combina diferentes redes para maximizar o aprendizado paralelo, conhecido como algoritmo de árvore de decisão de votação paralela. Devido à sua eficiência, precisão e interpretabilidade, o GBDT alcança alto desempenho em muitas tarefas de aprendizado de máquina, como classificação multiclasse e predição.

Uma característica distintiva do *LightGBM* é que ele divide a árvore folha por folha, ao contrário de outros algoritmos de GBDT que desenvolvem a árvore por nível. Ele seleciona a folha com uma perda máxima para crescer, onde a função de perda é uma medida de ajuste dos coeficientes do modelo e é utilizada para que nas etapas do reforço de gradiente os erros sejam minimizados. Assim, a estrutura continua a crescer com os ramos e folhas mais promissores, isto é, nós com a maior perda, mantendo constante o número de folhas de decisão. Em conjuntos de dados limitados, o desenvolvimento em que se utiliza o crescimento por folha pode aumentar a complexidade do modelo e levar ao sobreajuste. O *LightGBM* introduz algumas otimizações que o tornam mais rápido e eficiente:

- *Leaf-wise Growth*: em vez de crescer a árvore de decisão horizontalmente (nível por nível), o *LightGBM* cresce a árvore verticalmente, escolhendo o nó que mais reduz a perda. Isso reduz significativamente o número de nós necessários para atingir o mesmo nível de profundidade, resultando em uma árvore mais profunda e menos ramificada;
- *Histogram-based Algorithm*: o *LightGBM* utiliza um algoritmo baseado em histograma para encontrar os melhores pontos de divisão nos dados. Isso reduz a complexidade computacional, especialmente em conjuntos de dados grandes, tornando o treinamento mais rápido;
- *Gradient-based One-Side Sampling*: durante o treinamento, o *LightGBM* usa um método de amostragem baseado em gradiente para selecionar os exemplos de treinamento que mais contribuem para a redução da perda. Isso ajuda a focar nos exemplos mais importantes e acelerar o processo de treinamento.

Além disso, o *LightGBM* oferece suporte a várias técnicas de regularização, como a regularização L1 (Lasso) e L2 (*Ridge*), para evitar *overfitting* e melhorar a generalização do modelo. Ele também possui recursos avançados de otimização, como suporte a GPU e paralelismo, permitindo treinamento rápido em grandes conjuntos de dados.

LightGBM é amplamente utilizado em competições de ciência de dados e em aplicações da indústria devido à sua eficiência, escalabilidade e desempenho superior. Ele se tornou uma escolha popular para uma variedade de problemas de aprendizado de máquina, especialmente em conjuntos de dados grandes e complexos.

2.2 Trabalhos Relacionados

Essa seção aborda estudos relevantes e correlacionados a sete trabalho que aplicaram técnicas de aprendizado de máquina em contextos industriais para otimização de processos complexos. Esses estudos exploram desde a predição de alvura em polpa de celulose, a previsão de vazão em sistemas de separação bifásica na indústria de petróleo, a classificação de tensões em chapas de Aço IF utilizando sinais de correntes parasitas pulsadas e até a predição do incremento médio anual volumétrico de Eucalyptus.

2.2.1 Sensor virtual de alvura em polpa branqueada de celulose baseado em Inteligência Artificial

Paula (2022) investigou a aplicação de técnicas de AM para a predição da alvura da polpa de celulose, visando a otimização do processo de branqueamento na indústria de celulose. O estudo propôs a utilização Aprendizado de Máquina para predição da alvura da polpa de celulose. A escolha dos atributos relevantes para a criação do modelo de predição, bem como o período de histórico em estudo, foi realizada em conjunto com operadores e engenheiros de processo, sendo posteriormente validada com o uso de algoritmos computacionais estatísticos.

Para construção dos modelos preditivos foram utilizados diferentes tipos de algoritmos baseados em árvores de decisão, avaliados e comparados a acurácia obtida e outros indicadores de desempenho entre eles. Foram comparados os algoritmos *Decision Tree Regressor*, *Random Forest*, *XGBoost* e *LightGBM*.

Os resultados mostraram que, mesmo após a etapa de ajuste de sintonia de hiperparâmetros, o *LightGBM* e o *XGBoost* apresentaram os melhores desempenhos. A melhor combinação de hiperparâmetros para o *XGBoost* foi: *n_estimators* = 570, *max_depth* = 5 e *learning_rate* = 0,05, resultando em um MAE de 0,19818, MSE de 0,07030 e R^2 de 0,95222. Para o *LightGBM*, a melhor combinação de hiperparâmetros foi: *num_leaves* = 20, *learning_rate* = 0,15, *n_estimators* = 248 e *max_depth* = 5, resultando em um MAE de 0,19752, MSE de 0,07009 e R^2 de 0,95237.

Após a etapa de validação cruzada para garantir a generalização dos modelos em dados não conhecidos, evitando o *overfitting*, observou-se um decréscimo no desempenho dos modelos em relação aos dados de treinamento, como esperado. Ainda assim, tanto o *XGBoost* quanto o *LightGBM* conseguiram atingir as metas de acurácia estabelecidas, com R^2 superior a 80%. Na validação cruzada com 10 folds, o *XGBoost* apresentou um R^2 de 0,93784, representando uma queda de apenas 1,5%. Já o *LightGBM*, também na validação cruzada com 10 folds, alcançou um R^2 de 0,93913, uma queda de cerca de 1,4%. Já os modelos baseados em *Random Forest* e *Decision Tree* não alcançaram essas metas, tornando-se inviáveis para uso.

Os autores do trabalho concluíram que a viabilidade de aplicar esses modelos em outros processos da indústria de papel e celulose depende da execução adequada das etapas de pré-processamento, extração de padrões e pós-processamento na construção do modelo. O autor também sugere a utilização de novos algoritmos de aprendizado de máquina, como redes neurais artificiais, para aumentar a acurácia e poder de generalização do modelo.

2.2.2 Instrumento Virtual Inteligente para Previsão de Emulsão Água/Óleo

Campos (2022) aborda a aplicação de algoritmos de aprendizado de máquina para prever a vazão de líquido em um separador bifásico de uma plataforma de processamento de petróleo. Na indústria do petróleo, a medição precisa da vazão dos fluidos produzidos é crucial para a correta mensuração dos volumes de petróleo, água e gás, bem como para a otimização do sistema, redução de perdas de produção e aumento da eficiência do processo.

Para alcançar esses objetivos, Campos (2022) desenvolveu uma base de dados histórica das variáveis do separador bifásico, a qual foi utilizada para treinar e validar diversos algoritmos de aprendizado de máquina, incluindo *Random Forest*, Support Vector Machines (SVM), k-Nearest Neighbor (KNN) e *XGBoost*. A avaliação dos modelos foi realizada utilizando métricas de regressão amplamente aceitas, como raiz do erro quadrático médio (Root Mean Square Error (RMSE)), erro percentual absoluto médio (MAPE) e R^2.

Os resultados demonstraram a eficácia dos modelos, com o *Random Forest* apresentando o melhor desempenho (RMSE = 57.75, MAPE = 2.2, R^2 = 0.88). O *XGBoost* e o KNN também mostraram resultados sólidos (RMSE = 59.45 e 61.88, MAPE = 2.3 e 2.39, R^2 = 0.87 e 0.86, respectivamente), enquanto o modelo SVM teve o desempenho mais baixo (RMSE = 66.83, MAPE = 3.6, R^2 = 0.84).

Este estudo evidenciou o potencial do aprendizado de máquina para melhorar os processos industriais no setor petrolífero. Há uma correlação entre o trabalho de Campos (2022) e o presente estudo, uma vez que ambos utilizam métodos de aprendizado de máquina para prever variáveis críticas em processos industriais. Em ambos os estudos, o *XGBoost* foi um dos algoritmos aplicados, demonstrando sua robustez e eficácia em cenários de predição complexos.

2.2.3 Classificação de tensões em chapas de Aço IF utilizando aprendizado de máquina aplicado a sinais de correntes parasitas pulsadas

No trabalho de Larocca *et al.* (2022), o processo de treino, teste e validação foi realizado utilizando técnicas de aprendizado de máquina para a classificação de tensões em chapas de aço IF, empregando classificadores SVM e *LightGBM*. A divisão exata dos dados entre treino e teste, assim como a possível realização de validação cruzada, não foi explicitamente mencionada. Não há informações claras sobre a replicação dos experimentos para garantir a robustez dos resultados.

A configuração dos experimentos envolveu a aquisição de sinais de Correntes Parasitas Pulsadas (Pulsed Eddy Current (PEC)) das amostras e a aplicação de diferentes técnicas de pré-processamento. Sem processamento adicional, os sinais PEC resultaram em 30,85% de acurácia e 17,98% de F1-score. Após a aplicação da Principal Component Analysis (PCA), a acurácia média aumentou para 40,29% e o F1-score para 35,87%. A aplicação da Linear Discriminant Analysis (LDA) aos sinais PEC aumentou a acurácia para 72,10% e o F1-score para 69,95%. Com a aplicação da Discrete Fourier Transform (DFT) aos sinais de entrada, a acurácia atingiu 73,42% e o F1-score 70,28%. No entanto, ao combinar DFT e PCA, houve uma redução na acurácia para 67,96% e no F1-score para 63,89%. Em contraste, a combinação de DFT e LDA resultou na maior acurácia e F1-score para o *LightGBM*, com 85,28% e 84,91%, respectivamente.

Comparando os dois modelos, o SVM apresentou desempenho superior ao *LightGBM* nos conjuntos de dados PEC, PEC + PCA, PEC + LDA e DFT + LDA. O *LightGBM* destacou-se no pré-processamento DFT. Para o melhor caso, a combinação de DFT e LDA permitiu que o SVM alcançasse 86,10% de acurácia e 86,08% de F1-score, enquanto o *LightGBM* atingiu 85,28% de acurácia e 84,91% de F1-score.

No estudo de Larocca *et al.* (2022), o uso do *LightGBM* em conjunto com técnicas como a PCA, LDA, e DFT demonstrou melhorias significativas na acurácia e no F1-score, evidenciando a eficácia do pré-processamento de dados e a sintonia de hiperparâmetros. Também ficou demonstrado que a validação cruzada e a sintonia de hiperparâmetros foram práticas essenciais para garantir a generalização dos modelos para dados não conhecidos.

2.2.4 Predição do Incremento Médio Anual Volumétrico de Eucalyptus com Aprendizado de Máquina

Lopes *et al.* (2023) aplicou algoritmos de AM para predição futura do Incremento Médio Anual Volumétrico (IMAVol) (m³/ha/ano) de eucalipto. O conjunto de dados utilizado é composto de variáveis fisiológicas e o IMAVol de plantas de eucalipto de um projeto de melhoramento genético florestal.

O processo de treino, teste e validação foi realizado utilizando a técnica de validação cruzada k-fold. Esse procedimento é repetido até que todas as partições tenham sido usadas como treinamento e validação ao menos uma vez. Os dados foram divididos em 10 partições (*k-fold* igual a 10) e o processo de validação cruzada foi repetido 50 vezes para garantir uma avaliação robusta. Ao final das iterações, a média dos 50 resultados para o RMSE e R^2 foi calculada e considerada como resultado final para cada métrica.

A melhor performance foi obtida pelo algoritmo Random Forestr (RF) ao utilizar dados de 6, 18, 30 e 36 meses de idade das plantas. Com uma média de 2,84 ± 0,02 e 0,83 ± 0,03 para as métricas de RMSE) e R^2, respectivamente, apos a realização das 50 iterações. O *XGBoost* apresentou um desempenho competitivo, especialmente quando utilizado em combinação com as idades de 6, 18 e 30 meses. Esses resultados foram considerados promissores pelos autores para apoiar a seleção precoce de materiais genéticos de alta produtividade volumétrica.

2.3 Considerações Finais

Este capítulo proporcionou uma compreensão profunda dos processos essenciais na fabricação de polpa celulósica. Inicialmente, foi enfatizada a relevância econômica significativa da indústria de papel e celulose no Brasil, evidenciando seu impacto substancial tanto no faturamento quanto nas exportações do país. Apresentando dados atualizados do setor de árvores plantadas e sublinhando a importância estratégica da celulose dentro desse contexto.

A estrutura típica de uma fábrica de celulose foi descrita, delineando as funções essenciais da linha de fibras e da recuperação química. Especial atenção foi dada ao processo de cozimento contínuo *kraft* de fibras curtas e ao branqueamento de celulose ECF, incluindo os agentes branqueadores utilizados e os parâmetros críticos de controle como o número *Kappa* e a viscosidade da polpa.

O trabalho de Paula (2022) é o único que compartilha o mesmo segmento industrial e proporcionou uma base valiosa ao explorar técnicas de aprendizado de máquina para problemas similares. Os trabalhos são complementares na aplicação direta e específica na indústria de celulose com o foco na predição de alvura como um indicador crítico de qualidade. No entanto, este estudo se avança ao implementar o modelo de forma online, com dados reais de processo, após as etapas de treino, teste e validação.

O trabalho de Campos (2022) e o presente estudo de predição de alvura se correlacionam uma vez que ambos utilizam métodos de aprendizado de máquina para prever variáveis críticas em processos industriais. Em ambos os estudos, o *XGBoost* foi um dos algoritmos aplicados, demonstrando sua robustez e eficácia em cenários de predição complexos. No caso da predição de alvura, o *XGBoost* foi usado para modelar a alvura da polpa de papel, considerando diversas variáveis do processo de branqueamento.

O trabalho de Larocca *et al.* (2022) se correlaciona com o presente trabalho de predição de alvura pois ambos utilizam o *LightGBM* e técnicas de redução de dimensionalidade para otimizar o desempenho dos modelos preditivos. Ainda que a aplicação das técnicas de redução de dimensionalidade não tenha implicado em aumento de robustez do modelo do presente trabalho.

O trabalho de Lopes *et al.* (2023) sobre a predição do Incremento Médio Anual Volumétrico (IMAVol) de eucalipto com algoritmos de aprendizado de máquina tem uma correlação com este trabalho, especialmente no uso do *XGBoost*. Ambos os trabalhos aplicam técnicas de aprendizado de máquina para prever variáveis críticas em seus respectivos domínios: o IMAVol na silvicultura e a alvura na indústria de papel. A consistência e eficácia do *XGBoost* em ambos os estudos sublinham sua capacidade de lidar com problemas complexos de regressão, fornecendo resultados precisos e confiáveis em diferentes contextos industriais e de pesquisa.

3 Metodologia

Este capítulo apresenta a metodologia adotada para a construção e avaliação dos modelos preditivos de alvura no processo de branqueamento. O desenvolvimento do trabalho foi realizado na unidade de Aracruz da empresa Suzano Papel e Celulose, líder global na produção de celulose de eucalipto. O foco está na aplicação de técnicas de regressão, incluindo a Regressão Linear, XGBoost e LightGBM, para prever a alvura de saída do estágio de branqueamento, utilizando uma coleção de dados extensa e variada.

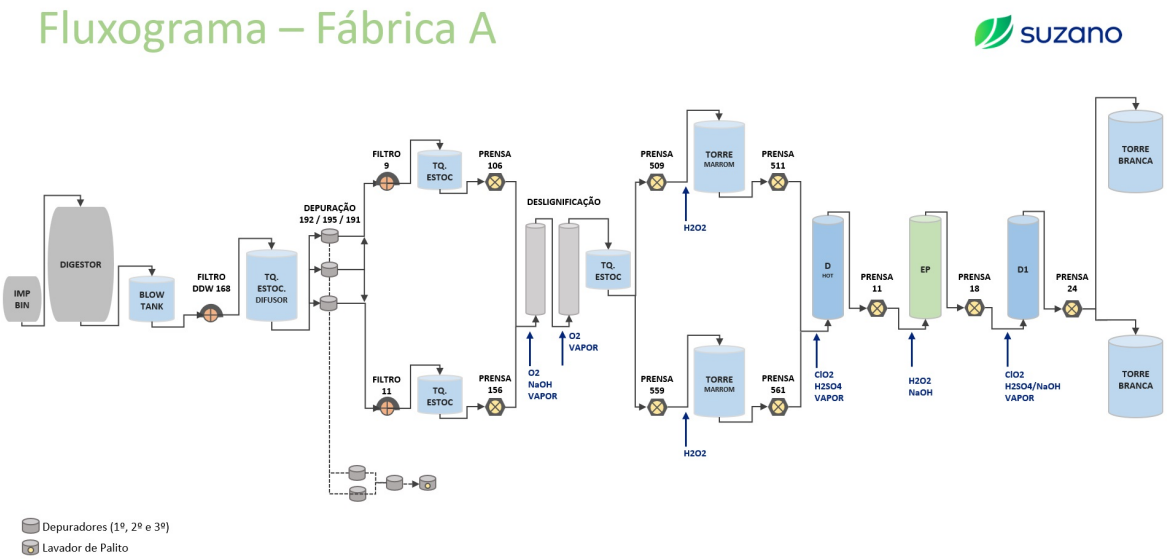
Nas seções deste capítulo, serão descritos ainda a coleção de dados utilizada e a seleção das variáveis. Em seguida, é detalhada a descrição dos experimentos de regressão realizados, incluindo os procedimentos para divisão dos dados em conjuntos de treino e teste, a aplicação da técnica de validação cruzada, e a otimização dos hiperparâmetros dos modelos. Foram adotadas estratégias como a divisão dos dados entre treinamento e teste na proporção de 80/20 e a implementação da técnica de validação cruzada com 5 *k-fold*. Esses passos metodológicos são essenciais para garantir a robustez e a precisão dos modelos desenvolvidos, permitindo uma previsão mais confiável da alvura de saída no processo de branqueamento. Inicialmente, é discutida a importância da compreensão do problema e da definição dos objetivos do estudo, destacando a relevância da alvura como atributo chave nesse contexto industrial. A coleção de dados utilizada foi descrita em detalhes, abrangendo as principais variáveis de entrada e a variável de saída, bem como as características e peculiaridades do conjunto de dados. Além disso, foram aplicadas técnicas de pré-processamento de dados, como a detecção e tratamento de valores ausentes, normalização e padronização das variáveis, e métodos de seleção de variáveis como o PCA.

3.1 Descrição da Empresa Envolvida

O trabalho foi desenvolvido na empresa Suzano Papel e Celulose, líder global na produção de celulose de eucalipto e uma das maiores fabricantes de papéis no mercado internacional. A produção anual da Suzano atinge expressivos 10,9 milhões de toneladas de celulose em sete unidades no Brasil. O projeto será desenvolvido na unidade de Aracruz-ES, responsável por 2,3 milhões deste volume. Essa unidade possui três linhas de produção, denominadas Fábrica A, B e C, esse trabalho foi desenvolvido na fábrica A, especificamente no Branqueamento A.

Na Figura 4 é demonstrado o fluxograma da linha de fibras da Fábrica A de Aracruz, onde será desenvolvido o trabalho de previsão de alvura.

Figura 4 – Fluxograma macro da linha de fibras da Fábrica A da unidade Suzano Aracruz.



Fonte: [O autor].

3.2 Descrição da Coleção de Dados e Caracterização dos Dados

A coleção de dados utilizada nesta pesquisa compreende uma lista de variáveis de processo, cada uma identificada por um TAG específico. A Tabela 2 apresenta uma síntese das principais variáveis, incluindo o TAG associado, a descrição do TAG, o tipo de variável e a faixa de valor.

Dentre as variáveis destacadas, englobam-se parâmetros importantes para o processo de branqueamento, como a produção do branqueamento A, alvura de entrada e saída do estágio, temperatura de entrada do estágio, *kappa* de entrada, pH de entrada, *shives* de entrada, carga de dióxido, teor de HexA, tipo de produto buscado no branqueamento e o fluxo de ácido sulfúrico.

Para garantir a integridade e confiabilidade dos dados, foram adotadas estratégias específicas no pré-processamento da coleção de dados. A seleção de intervalos de valores para permanecer na análise foi realizada considerando o conhecimento técnico de processo e a presença de extremos ou inconsistências que poderiam distorcer a análise.

Para isso, foram aplicados filtros específicos, como limites de valores aceitáveis, garantindo que apenas dados dentro de faixas pré-estabelecidas fossem incluídos na análise, conforme demonstrado na Tabela 4. O principal filtro ocorreu na variável “Produção do Branqueamento A”, no qual apenas foram considerados dados em que o branqueamento estivesse operando acima de 90.0% da produção nominal.

Essa abordagem é necessária pois em períodos de produção reduzida algumas variáveis são favorecidas e outras prejudicadas. Isso não reflete a realidade de processo, visto que a planta opera com *Operational Stability* (OS) em produção nominal superior a 91.0%.

Tabela 2 – Descrição de todas as variáveis extraídas do processo.

TAG	**Descrição do TAG**	**Tipo de variável**	**Faixa de Valor**
205F260.PV	Produção do Branqueamento A	Numérica	$[1400, 2000]$
205Q230B.PV	Alvura de entrada do Estágio DHT	Numérica	$[55, 65]$
205T031.PV	Temperatura de entrada do Estágio DHT	Numérica	$[75, 90]$
205Q230A.PV	Kappa de entrada do Estágio DHT	Numérica	$[9, 11]$
205Q033.PV	pH de entrada do Estágio DHT	Numérica	$[2.5, 4]$
205Q20224.PV	Shives de entrada do Estágio DHT	Numérica	$[0.1, 80]$
205Q027C.PV	Carga de dióxido Estágio DHT	Numérica	$[2, 10]$
205QHEX_A_101.PV	teor de HexA na entrada do estágio DHT	Numérica	$[60, 80]$
205Q230D.PV	Alvura de saída do Estágio DHT	Numérica	$[50, 77]$
205N008.PV	Consistência de entrada do Estágio DHT	Numérica	$[0, 15]$
205F027.PV	Fluxo dióxido Estágio DHT	Numérica	$[5, 90]$
205H250.PV	Tipo de Produto produzido	Categórica	[AXP-LB, AXP-LB2]
205F004.PV	Fluxo de ácido Estágio DHT	Numérica	$[0, 20]$

Fonte: elaborado pelo autor.

Nesta etapa também foram selecionadas as principais varáveis de entrada que seriam levadas para os modelos. As variáveis “Fluxo dióxido Estágio DHT” e “Consistência de entrada do Estágio DHT” não serão levadas para o modelo, pois serão utilizadas apenas no cálculo do tempo de retenção da torre DHT, a fim de deslocar no tempo as análises de alvura de saída em relação as variáveis de entrada.

A variável “pH de entrada do Estágio DHT” não foi considerada adequada para uso nos modelos de predição, visto que há um controle de processo que atua para que não ocorra variação de pH. Uma variável que apresenta pouca variação tem baixa relevância para medir a qualidade da polpa de entrada. Dessa forma, ela foi substituída pela variável “Fluxo de ácido Estágio DHT”, pois este controle atua para ajustar o pH variando o fluxo de ácido sulfúrico.

A variável "Tipo de Produto produzido" também não foi utilizada no controle, visto que há pouca diferença entre os produtos, em termos de alvura final. No estágio DHT, o primeiro do branqueamento A, os dois produtos almejam a mesma alvura de saída. Além disso, essa é uma variável "Categórica" e inserção dela resultou em menor assertividade do modelo. Por todos esses modelos ela não foi inserida entre as variáveis de entrada dos modelos.

A "Alvura de saída do Estágio DHT" é a variável de saída que se pretende prever no trabalho, portanto também não figura entre as variáveis de entrada entregues aos modelos.

Para todos as 8 variáveis restantes, foi realizada a tabela de sumarização estatística, conforme Tabela 3. Cada variável está descrita em termos de média, desvio padrão, valor mínimo, percentil de 25%, mediana (percentil de 50%), percentil de 75%, valor máximo e o Coeficiente de Variação (CV). Enquanto a média representa o valor médio das observações, o desvio padrão indica a dispersão dos dados em relação à média. Os valores mínimo e máximo mostram os extremos dos dados, indicando o intervalo de variação possível. Os percentis (25%, 50% e 75%) dividem os dados ordenados em quartis, proporcionando uma visão sobre a distribuição dos dados ao longo desses pontos. O CV mede a variabilidade relativa de uma distribuição de dados em relação à sua média.

Essas estatísticas fornecem uma visão das variáveis e ajudam a entender sua distribuição e variabilidade, além de caracterizar a coleção de dados, que conta com um total de 7978 registros. A partir disso, é possível identificar tendências centrais, dispersão e possíveis valores discrepantes em cada uma, dando subsidio para as análises subsequentes.

Tabela 3 – Descrição das Variáveis de processo.

Variável	**média**	**desvio padrão**	**min**	**25%**	**50%**	**75%**	**max**	**CV (%)**
Fluxo Acido	2.29	0.88	0.50	1.67	2.25	2.83	6.34	38.4
Producao branq	1674.4	105.7	1500.0	1599.8	1698.5	1766.6	1996.6	6.3
Carga dioxido	4.93	1.17	2.42	4.30	4.56	5.52	10.00	23.8
Shives	12.68	6.29	0.72	8.84	11.01	14.79	59.40	49.6
Kappa entrada DHT	10.15	0.50	8.57	9.69	10.28	10.56	11.39	4.9
Alvura entrada DHT	58.63	1.79	52.22	57.36	58.67	59.90	64.99	3.0
Teor Hexa	71.80	3.26	60.11	68.92	72.56	74.35	79.99	4.5
Temperatura entrada	87.15	3.21	75.47	85.73	87.88	89.39	94.30	3.7

Fonte: [o autor].

Para as mesmas variáveis de entrada do modelo, também foi realizada a Função de Distribuição Acumulada (Cumulative Distribution Function (CDF)), fornecendo uma representação detalhada da distribuição de probabilidade dessas variáveis. A CDF é uma função que descreve a probabilidade de uma variável aleatória ser menor ou igual a um determinado valor.

Ao calcular a CDF para cada variável de entrada, é possível visualizar como os dados estão distribuídos ao longo de todo o intervalo de valores possíveis. Isso permite uma melhor compreensão da variabilidade dos dados e ajuda na identificação de padrões ou comportamentos anômalos que podem afetar o desempenho do modelo preditivo.

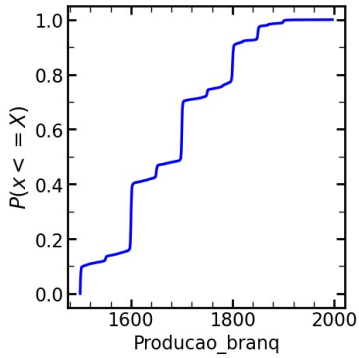
Na Figura 5a é apresentada a CDF da "produção do branqueamento" onde é possível observar que a planta opera em faixas de valores muito específicas, ou seja, há muitos valores em cima de 1600.0, 1700.0 e 1800.0 tSA/d. Isso ocorre porque esses são valores típicos de produção setados no branqueamento, a qual não é comum operar com valores fracionados ou intermediários de produção.

A CDF da "alvura de entrada DHT" é apresentada na Figura 5b, nessa figura é possível verificar que aproximadamente 60.0% dos dados estão entre 57.0 e 60.0 (%ISO). Esses valores são considerados muito bons, pois se tratando de uma planta com muitos equipamentos da década de 1970, ele ainda performa em patamares compatíveis com unidades mais novas. Da mesma forma, a CDF de "Kappa de entrada DHT" na Figura 5c mostra que aproximadamente 60.0% dos dados estão entre 9.5 e 10.5. Essa distribuição de valores é boa, pois indica uma variação de 0.5 em relação ao valor médio ideal de 10.0.

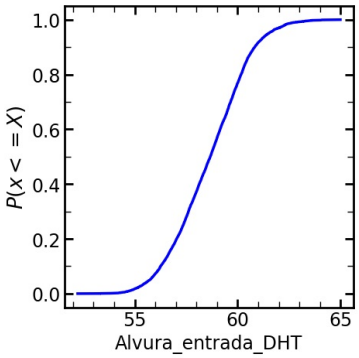
A partir da CDF de "Shives na entrada DHT" apresentada na Figura 5d é possível identificar que aproximadamente 90.0% dos dados são inferiores a 20.0 palitos/grama, valor de referência para não ocorrer desclassificação do produto por sujidade elevada. Por sua vez, a CDF do "hexa de entrada DHT" exibida na Figura 5e mostra que aproximadamente 80.0% dos dados estão entre 65.0 e 75.0. Esse resultado demonstra que o HexA compõe parte significativa do *kappa* na entrada do branqueamento.

Na Figura 5f é apresentada a CDF da "Carga Dióxido DHT" em que é possível verificar que aproximadamente 60.0% do tempo esse químico para o estágio fica entre 4.0 e 6.0 Kg/tSA. De semelhante modo, a partir da CDF do "Fluxo ácido DHT" na Figura 5g, é possível atestar que aproximadamente 70.0% do tempo a vazão desse químico fica entre 2.0 e 4.0 l/min para controlar o pH do estágio. Por fim, a CDF da "Temperatura Estágio DHT" na Figura 5h mostra que aproximadamente 80.0% do tempo a temperatura desse estágio fica acima de 85.0°C, valor considerado ideal pela literatura para a adequada hidrólise ácida de HexA.

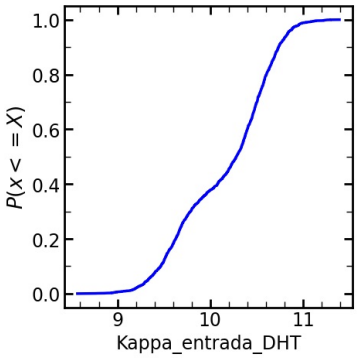
Figura 5 – Distribuição das variáveis: Função de Distribuição Acumulada.



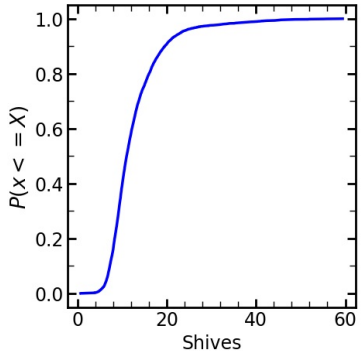
(a) Produção branqueamento.



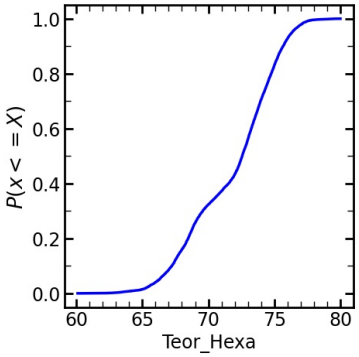
(b) Alvura entrada.



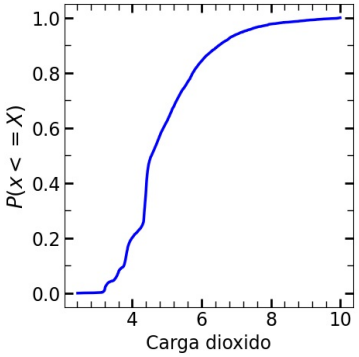
(c) Kappa entrada.



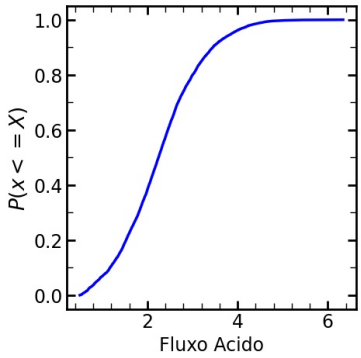
(d) Shives.



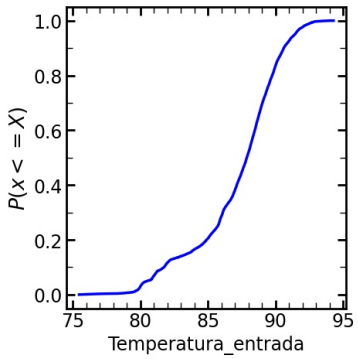
(e) Teor de hexa.



(f) Carga dióxido.



(g) Fluxo ácido.



(h) Temperatura de entrada.

Fonte: [O autor].

3.3 Importância das *Features* no Problema

Na etapa de seleção de variáveis (*features*) deste estudo, foi adotado um processo detalhado visando identificar as variáveis mais relevantes para a eficácia do modelo preditivo. No contexto da predição da alvura do estágio de branqueamento de polpa de celulose, as variáveis selecionadas desempenham papéis específicos na determinação do resultado final. Vamos examinar cada uma delas e sua importância técnica:

- **Produção do Branqueamento (*Producao_branq*)**: a produção de branqueamento é uma medida do volume de polpa de celulose processada em determinado período de tempo. Ela pode afetar significativamente a eficiência do processo e constitui o principal filtro utilizado, visto que determina quando a fábrica está "parada";
- **Alvura de Entrada do Estágio DHT (*Alvura_entrada_DHT*)**: a alvura de entrada do estágio é uma das principais características para a ser monitorada, visto que ela vai afetar fortemente a alvura de saída;
- **Kappa de Entrada do Estágio DHT (*Kappa_entrada_DHT*)**: o número kappa é uma medida da quantidade de lignina e HexA na polpa de celulose. O valor do kappa afeta diretamente o processo de branqueamento, pois a lignina é o principal grupo cromóforo (que apresenta cor) da polpa;
- **Teor de Shives (*Shives*)**: os *Shives* são pedaços incozidos de madeira que podem estar presentes na polpa de celulose em quantidade, por problemas na etapa de depuração. Sua presença afeta a qualidade do produto final e, portanto, monitorar a quantidade de *Shives* é essencial para garantir que ocorra a dosagem adequada de químicos para controlá-lo;
- **Teor de HexA na entrada do estágio DHT (*Teor_Hexa*)**: o teor de hexa é uma medida da quantidade de ácidos hexenuronicos na polpa de celulose. Os HexA's são formados pela descarboxilação dos grupos de ácido glucurônico presentes nas xilanas durante a polpação alcalina (cozimento da madeira);
- **Carga de Dióxido de Cloro (*Carga_dioxido*)**: o dióxido de cloro é um dos agentes de branqueamento mais comuns na indústria de papel e celulose. A carga de dióxido de cloro influencia diretamente a eficácia do branqueamento, por se tratar do principal químico utilizado para branquear a polpa;
- **Fluxo de Ácido (*Fluxo_Acido*)**: essa variável quantifica o volume de ácido sulfúrico necessário para controlar o pH do estágio. Nesse modelo, optou-se por utilizar o fluxo de ácido no lugar da variável pH, visto que o controle age para que o pH sempre fique constante e aderente ao *setpoint*. Para um modelo que busca correlacionar a variação das *features* de entradas com a alvura de saída, ela não agregaria;
- **Temperatura de Entrada (*Temperatura_entrada*)**: A temperatura de entrada afeta a velocidade das reações químicas no processo de branqueamento. Uma temperatura adequada é essencial para garantir a eficácia do processo e a qualidade do produto final.

A seleção cuidadosa dessas variáveis e seu uso técnico na modelagem garantem que o modelo seja capaz de fornecer previsões precisas e úteis, auxiliando na otimização do processo de branqueamento e na melhoria da qualidade do produto final.

3.4 Análise de Correlação e Seleção de *Features*

Na etapa de pré-processamento foram realizados filtros de valores em cada variável de entrada do modelo de predição. Essas faixas de valores estão descritas na Tabela 2. Esse procedimento consiste em remover os inconsistentes ou *outliers* que estão fora dos patamares normais de operação, para que eles não distorçam a análise e prejudiquem a qualidade das previsões geradas pelo modelo.

Neste trabalho foram coletadas 13 variáveis típicas do processo de branqueamento de celulose, mas após as remoções citadas na Seção 3.3, as 8 variáveis restantes foram submetidas a análise da correlação de *Pearson*. Essa medida estatística avalia a relação linear entre duas variáveis contínuas.

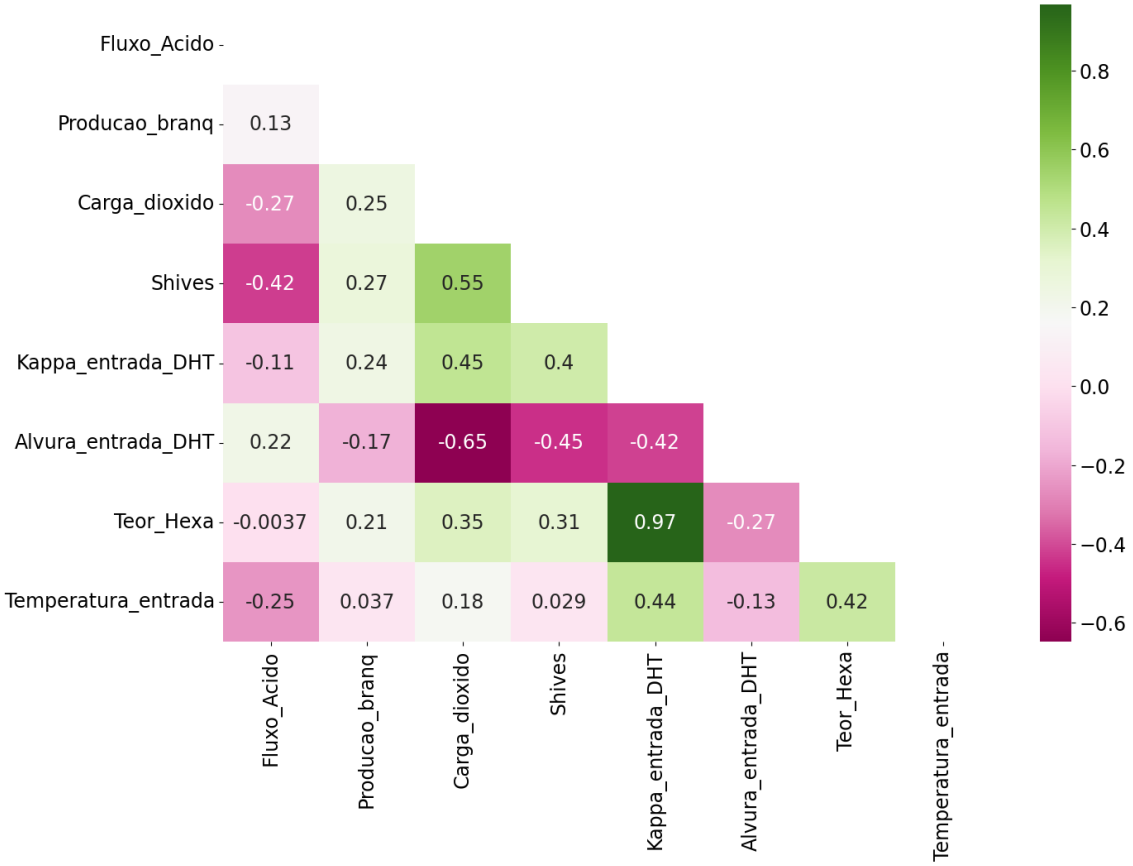
Com a métrica de *Pearson*, as variáveis "Kappa de entrada do Estágio DHT" e "teor de HexA na entrada do estágio DHT" apresentaram uma correlação de 0.97, considerada correlação alta. Esse resultado está em conformidade com a literatura, visto que o teor de HexA compõe o que é medido na análise de kappa. Devido a isso, foi realizada a exclusão da variável "teor de HexA na entrada do estágio DHT" do modelo preditivo.

Além disso, dois outros pares de variáveis apresentaram correlações moderadas (entre 0.5 e 0.75). A correlação entre "*Shives* de entrada do Estágio DHT" e "Carga de dióxido Estágio DHT" foi de 0.55, o que indica uma relação positiva moderada, sugerindo que um aumento na quantidade de *Shives* está associado a um aumento na quantidade de dióxido de cloro utilizado. Isso pode ser explicado pelo fato de que maiores quantidades de *Shives*, que são partículas de madeira não completamente processadas, exigem mais agente branqueador para atingir os níveis desejados de alvura.

Por outro lado, a correlação entre "Alvura de entrada do Estágio DHT" e a "Carga de dióxido Estágio DHT" foi de -0.65, o que indica uma relação negativa moderada e sugere que uma maior alvura inicial da polpa está associada a uma menor necessidade de dióxido de cloro. Ou seja, quando a alvura inicial é maior, menos agente branqueador é necessário para atingir a alvura final desejada, resultando em uma operação mais econômica e eficiente.

Na Figura 6 é apresentada o resultado da análise de correlação de *Pearson* para as 8 variáveis.

Figura 6 – Correlação de *Pearson* dos dados de entrada do modelo.



Fonte: [O autor].

A manutenção de variáveis de alta correlação poderia levar o modelo a problemas de multicolinearidade, o que pode prejudicar na sua interpretação e levar a estimativas imprecisas dos coeficientes. Portanto, ao identificar e remover variáveis altamente correlacionadas, foi possível reduzir o "sobrepeso" de determinadas características nos dados e melhorar a eficiência do modelo.

A partir da Tabela 2 e considerando as premissas descritas no Capítulo 3 restaram as 7 variáveis de entrada utilizadas nos modelos conforme apresentados na Tabela 4.

Conforme informado anteriormente, as variáveis "Fluxo dióxido Estágio DHT" e "Consistência de entrada do Estágio DHT" foram adicionadas apenas para realização do cálculo do tempo de retenção da torre DHT, para este cálculo também foi utilizada a variável "Produção do Branqueamento A". Com isso, será possível determinar o tempo necessário para que a polpa passe pelo estágio, possibilitando o deslocamento no tempo das variáveis de entrada com a saída e a comparação da alvura real e a predita pelo modelo.

No cálculo do tempo de retenção da torre DHT, a fórmula utilizada é dada por:

$$\text{Tempo de Retenção} = \frac{1384}{\left(\text{Producao\_branq} \times \frac{90}{\text{Consistência\_entrada}} + \frac{\text{Fluxo\_dioxido}}{24}\right)} \times 24,$$

Tabela 4 – Descrição das variáveis de entrada selecionadas para construção dos modelos.

TAG	**Descrição do TAG**	**Tipo de variável**	**Faixa de Valor**
205F260.PV	Produção do Branqueamento A	Numérica	$[1400, 2000]$
205Q230B.PV	Alvura de entrada do Estágio DHT	Numérica	$[55, 65]$
205T031.PV	Temperatura de entrada do Estágio DHT	Numérica	$[75, 90]$
205Q230A.PV	Kappa de entrada do Estágio DHT	Numérica	$[9, 11]$
205Q20224.PV	*Shives* de entrada do Estágio DHT	Numérica	$[0.1, 80]$
205Q027C.PV	Carga de dióxido Estágio DHT	Numérica	$[2, 10]$
205F004.PV	Fluxo de ácido Estágio DHT	Numérica	$[0, 20]$

Fonte: [o autor].

em que:

- Producao_branq (tSA/d): ritmo de produção do branqueamento A;
- Fluxo_dioxido (m3/h): fluxo de dióxido de cloro na entrada da torre DHT;
- Consistência_entrada (%) : consistência da polpa na entrada da torre DHT;
- $1384(m3)$: volume da torre DHT;
- $90(\%)$: consistência do produto final;
- $24(h/d)$: fator de conversão entre dias e horas.

Ainda na etapa de pré-processamento, foi aplicado o Teste de Máximo e Mínimo nas variáveis de entrada do modelo, apresentado na Tabela 5. Nesse teste, o valor máximo da variável que opera com os maiores valores é dividido pelo valor mínimo da variável que opera com os menores valores. Isso é utilizado para verificar a presença de *outliers* ou de variáveis que performam em patamares muito diferentes.

Isso é feito pois variáveis com valores naturalmente maiores podem imprimir maior peso nos modelos de predição do que variáveis com valores menores. Esses valores muito discrepantes nas variáveis de entrada poderiam influenciar negativamente na performance dos algoritmos de aprendizado de máquina.

Dessa forma, esse procedimento permite avaliar a necessidade de aplicar a transformação logarítmica antes de prosseguir com a modelagem. Neste caso, foi necessária a aplicação, visto que o valor Máximo da variável "Produção do branqueamento" dividido pelo valor Mínimo da variável "Fluxo de Ácido" foi 3981.6. Esse valor indica que uma variável opera em patamar mais de três ordens de grandeza maior que a outra.

Tabela 5 – Valores Mínimos e Máximos das Variáveis.

Variável	**Mínimo**	**Máximo**
Produção do Branqueamento	1500.0	1996.6
Temperatura de Entrada	75.47	94.30
Alvura de Entrada DHT	52.23	64.99
Kappa de Entrada DHT	8.57	11.39
Carga de Dióxido	2.42	10.00
Teor de *Shives*	0.72	59.40
Fluxo de Ácido	0.501	6.340
Maior Valor / Menor Valor	**3981.61**	

Fonte: [o autor].

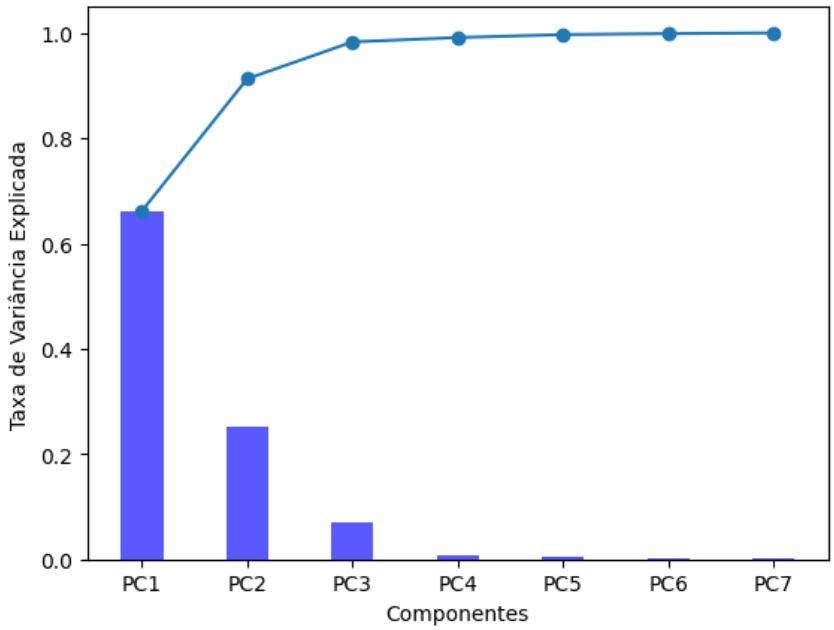
Com o resultado do teste de máximo e mínimo foi confirmada a necessidade de aplicação da Transformação logarítmica dos dados, pois o "Maior Valor / Menor Valor" maior do que 3 ordens de grandeza. Assim, nessa etapa foi aplicado a transformação logarítmica na base 10 em todas as variáveis de entrada e saída do modelo, dessa forma os números ficam mais próximos, reduzindo a discrepância e não imprimindo peso adicional a nenhuma variável.

A última etapa de pré-processamento de dados foi a Análise de Componentes Principais (PCA), que é uma técnica para reduzir a dimensão de um conjunto de dados, preservando suas propriedades.

O PCA é um método de seleção de características que auxilia na determinação das variáveis importantes em um modelo. Ou seja, ele ajuda a encontrar um modelo mais simples, com menos variáveis ou entradas, sem comprometer a qualidade do ajuste aos dados de treinamento. Ao reduzir a complexidade do modelo com o PCA, pode-se esperar que o modelo diminua o *overfitting* ao ser usado para previsão.

O resultados das componentes principais do PCA está demonstrado na Figura 7:

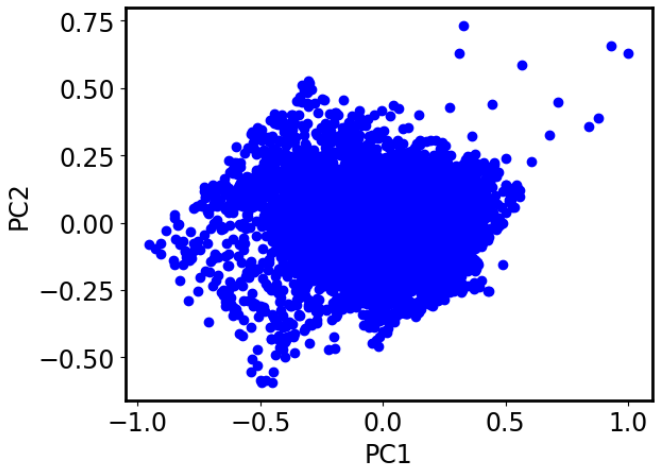
Figura 7 – Taxa de Variação Explicada de todos os Componentes Principais.



Fonte: [O autor].

Com o resultado do PCA é possível plotar O gráfico de dispersão das componentes principais. Ele demonstra a distribuição dos pontos de dados ao longo das duas componentes principais, após a transformação PCA. Ao examinar esse *scatterplot*, é possível identificar padrões, agrupamentos ou separações entre os dados, fornecendo informações sobre a estrutura subjacente dos dados, conforme representado na Figura 8.

Figura 8 – Gráfico de dispersão das componentes principais.



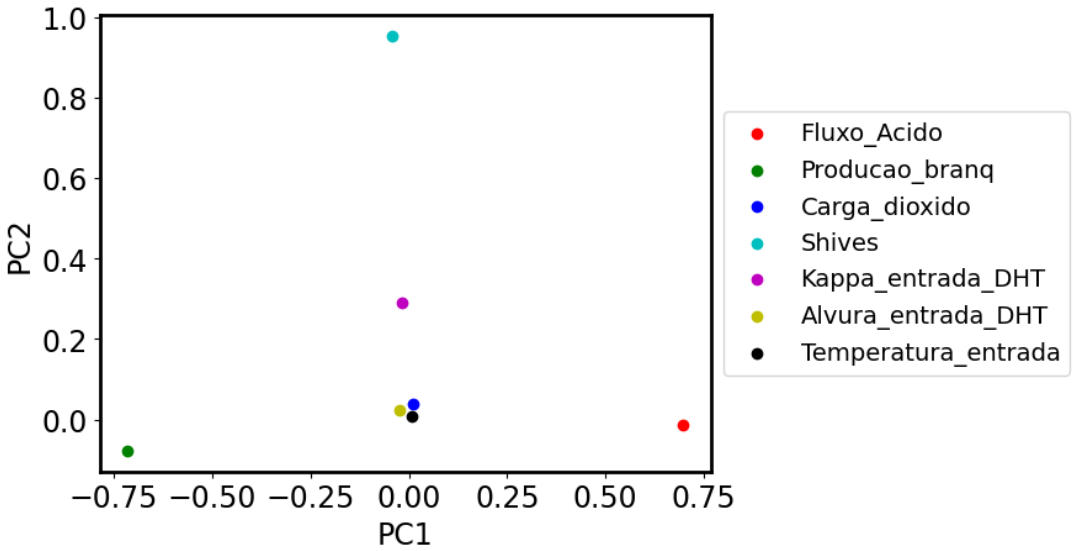
Fonte: [O autor].

A partir do resultado do PCA, também foi feito o gráfico de dispersão ou *scatterplot* das variáveis que impactam na PC1 e PC2. Esse gráfico permite visualizar como as variáveis originais contribuem para a variabilidade dos dois componentes principais, que capturam a maior parte da informação dos dados.

A partir dessa análise, é possível observar padrões de agrupamento, identificar possíveis correlações entre as variáveis e entender a importância de cada variável na formação das componentes principais. Um exemplo de percepção que é possível de extrair desse gráfico é de que variáveis que estão próximas umas das outras tendem a ter uma relação mais forte entre si, enquanto variáveis distantes tendem a ser mais independentes.

Na Figura 9 é apresentado o gráfico de dispersão das variáveis que impactam nos PC1 e PC2.

Figura 9 – Gráfico de dispersão dos coeficientes dos autovetores das variáveis que impactam no PC1 e PC2.



Fonte: [O autor].

Para compreender as variáveis que mais impactam no PC1, é necessário visualizar graficamente os pontos que ficaram mais distantes do zero no eixo das abscissas (x). Para compreender as que mais impactam no PC2, deve-se fazer o mesmo só que no eixo das ordenadas (y). Assim, analisando a Figura 9 é possível inferir:

- **Componente 1 (PC1)**: é mais impactado pelas variáveis “Fluxo_Acido” e “Producao_branq”.
- **Componente 2 (PC2)**: é mais impactado pelas variáveis “Shives” e “Kappa_entrada_DHT”.

O PCA é uma técnica amplamente utilizada para a redução de dimensionalidade, particularmente eficaz quando se lida com um grande número de variáveis correlacionadas. Ao transformar variáveis originais em um conjunto menor de componentes principais, o PCA visa capturar a maior parte da variância presente nos dados com menos dimensões. No entanto, em nosso caso específico, o número de variáveis originais (sete) não foi suficientemente grande para que a redução de dimensionalidade se mostrasse benéfica.

Quando os modelos foram treinados utilizando os PC's – PC1, PC2 e PC3 – não conseguiram superar a performance dos modelos treinados com as variáveis originais. Uma possível explicação para isso é que, com apenas sete variáveis, a complexidade e correlação entre elas podem ser gerenciadas diretamente pelo modelo preditivo sem a necessidade de transformação. Além disso, a transformação das variáveis originais em componentes principais pode ter levado à perda de informações relevantes que são capturadas mais efetivamente por elas.

Portanto, decidimos seguir com as variáveis originais, reconhecendo que a redução de dimensionalidade, embora útil em muitos contextos, não foi necessária ou benéfica neste caso específico. A escolha de fornecer as sete variáveis originais ao modelo foi baseada na observação empírica de um desempenho superior, destacando a importância de avaliar a eficácia das técnicas de pré-processamento de dados em função do problema e do conjunto de dados específico em análise.

3.5 Métricas de Avaliação dos Modelos

Para avaliar o desempenho dos modelos preditivos de alvura no processo de branqueamento, foram utilizadas três métricas principais: Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE) e o coeficiente de determinação (R^2). Cada uma dessas métricas oferece uma perspectiva diferente sobre a precisão e a capacidade de generalização dos modelos. A seguir, são apresentadas as definições matemáticas de cada métrica, suas interpretações e a importância de utilizá-las de forma complementar.

O Erro Médio Absoluto (MAE) é uma métrica que mede a média dos erros absolutos entre as previsões do modelo e os valores observados. Ele é calculado pela fórmula:

- **Mean Absolute Error (MAE)**:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|\,, \tag{3.1}$$

  em que:

  - n é o número total de observações;
  - y_i é o valor observado;
  - $\hat{y}_i$ é o valor predito.

O MAE captura a magnitude média dos erros sem considerar a direção dos mesmos (erro positivo ou negativo). Essa métrica é útil porque é intuitiva e fácil de interpretar, fornecendo uma medida clara de quão longe, em média, as previsões estão dos valores reais. No entanto, o MAE não penaliza grandes erros tanto quanto o MSE, o que pode ser uma limitação quando se deseja dar maior importância a grandes discrepâncias.

O Erro Quadrático Médio (MSE) mede a média dos erros ao quadrado entre as previsões do modelo e os valores observados. A fórmula é:

- **Mean Squared Error (MSE)**:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{3.2}$$

em que:

  - n é o número total de observações;
  - y_i é o valor observado;
  - $\hat{y}_i$ é o valor predito.

O MSE é sensível a grandes erros devido ao fato de elevar os erros ao quadrado. Isso significa que erros maiores têm um impacto significativamente maior no valor do MSE. Essa característica torna o MSE útil para situações em que grandes erros são particularmente indesejáveis. No entanto, o MSE pode ser influenciado por *outliers*, tornando-se menos representativo em conjuntos de dados com valores anômalos extremos.

O coeficiente de determinação (R^2) é uma métrica que indica a proporção da variabilidade nos dados que é explicada pelo modelo. A fórmula é:

- **Coeficiente de Determinação (R^2)**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.3}$$

em que:

  - n é o número total de observações;
  - y_i é o valor observado;
  - $\hat{y}_i$ é o valor predito;
  - $\bar{y}$ é a média dos valores observados.

O R^2 varia entre 0 e 1, onde 0 indica que o modelo não explica nenhuma variabilidade dos dados, enquanto 1 indica que o modelo explica toda a variabilidade dos dados. O R^2 é útil porque fornece uma medida relativa do quão bem o modelo se ajusta aos dados em comparação com um modelo simples que usa a média dos valores observados como previsão. No entanto, o R^2 pode ser enganoso em modelos não lineares ou quando se compara modelos com diferentes números de variáveis preditoras.

3.6 Descrição dos Experimentos de Regressão

O presente estudo utilizou diversas ferramentas tecnológicas avançadas para a predição de alvura da polpa. A principal linguagem de programação utilizada foi o *Python*, devido à sua ampla adoção na comunidade científica e sua vasta gama de bibliotecas específicas para aprendizado de máquina e análise de dados. O ambiente de execução escolhido foi o *Databricks*[1], uma plataforma de análise de dados baseada na nuvem, que facilita a colaboração e a escalabilidade dos experimentos.

Diversos pacotes *Python* foram empregados ao longo dos experimentos, cada um com suas funcionalidades específicas:

- **Pandas**: utilizado para manipulação e análise de dados, o Pandas[2] permite operações eficientes em grandes conjuntos de dados, proporcionando estruturas de dados de alto desempenho e ferramentas de análise que facilitam a limpeza, transformação e agregação de dados;
- **Scikit-learn (sklearn)**: essa biblioteca[3] foi fundamental para a implementação de algoritmos de aprendizado de máquina, incluindo a separação dos dados em conjuntos de treino e teste, a construção e avaliação de modelos, e a validação cruzada. O Scikit-learn oferece uma ampla gama de algoritmos de aprendizado de máquina, ferramentas para seleção de modelos e métricas de avaliação;
- **Matplotlib**: utilizada para visualização de dados, a Matplotlib[4] é uma biblioteca que permite a criação de gráficos estáticos, animados e interativos em Python. Foi essencial para a visualização inicial dos dados e para a plotagem de resultados dos modelos de predição;

[1] Databricks: <https://www.databricks.com/>.
[2] Pandas: <https://pandas.pydata.org/>.
[3] Scikit-learn: <https://scikit-learn.org/>.
[4] Matplotlib: <https://matplotlib.org/>.

- **Seaborn**: complementando o Matplotlib, o Seaborn[5] oferece uma interface de alto nível para a criação de gráficos estatísticos atraentes e informativos. Ele foi utilizado para a visualização de distribuições de dados e relações entre variáveis, facilitando a compreensão dos padrões e correlações nos dados;

- **Plotly**: essa biblioteca[6] foi utilizada para criar gráficos interativos, incluindo matrizes de confusão detalhadas e visualizações tridimensionais. O Plotly permite a criação de visualizações complexas e interativas, que são úteis para apresentações e análises detalhadas dos resultados;

- **NumPy**: fundamental para operações matemáticas e manipulação de arrays, o NumPy[7] fornece suporte para grandes matrizes multidimensionais e uma coleção de funções matemáticas de alto nível. Foi amplamente utilizado para cálculos numéricos e operações vetoriais necessárias nos algoritmos de aprendizado de máquina;

A combinação dessas ferramentas permitiu a execução eficiente e precisa dos experimentos, desde a preparação e análise dos dados até a construção, avaliação e visualização dos modelos preditivos. A escolha do *Databricks* como plataforma de execução proporcionou um ambiente escalável e colaborativo, essencial para o processamento de grandes volumes de dados e para a integração com o sistema de gerencialmente utilizado na fábrica atualmente, o PI System.

Os experimentos realizados neste estudo envolveram a aplicação de técnicas de regressão e classificação para prever a alvura de saída de um estágio de branqueamento de celulose. O primeiro modelo utilizado foi a regressão linear, que é uma abordagem básica amplamente empregada para modelagem preditiva. Este modelo assume uma relação linear entre as variáveis de entrada e a variável de saída, neste caso, a alvura de saída.

O segundo modelo adotado foi o *XGBoost*, algoritmo que se destaca pela sua eficiência e capacidade de lidar com conjuntos de dados extensos, oferecendo recursos avançados de regularização e possibilitando a construção de modelos altamente precisos. No contexto da avaliação do modelo, foi realizada uma divisão dos dados entre treinamento e teste na proporção de 80/20, garantindo uma separação apropriada para o treinamento e a validação dos modelos.

Além disso, foi implementada a técnica de validação cruzada com 5 *k-folds*, onde os dados foram divididos em cinco subconjuntos distintos para treinamento e validação iterativa dos modelos. Durante a validação cruzada, os hiperparâmetros de cada modelo foram ajustados e otimizados, com variação sistemática para encontrar as configurações mais adequadas. Essa abordagem permitiu uma avaliação abrangente do desempenho dos modelos e contribuiu para a seleção das melhores configurações para a previsão da alvura de saída no estágio de branqueamento.

[5] Seaborn: <https://seaborn.pydata.org/>.
[6] Plotly: <https://plotly.com/python/>.
[7] NumPy: <https://numpy.org/>.

Os principais hiperparâmetros do *XGBoost* ajustados nos experimentos foram:

- **Learning Rate**: controla a taxa de aprendizado do modelo. Valores menores tornam o treinamento mais robusto, mas podem exigir um número maior de iterações para convergir;
- **Max Depth**: define a profundidade máxima das árvores individuais. Árvores mais profundas podem capturar relações mais complexas, mas também aumentam o risco de *overfitting*;
- **N Estimators**: especifica o número de árvores a serem construídas. Um número maior de árvores geralmente melhora a precisão do modelo, mas também aumenta o tempo de treinamento;
- **Colsample Bytree**: representa a fração de colunas a serem amostradas aleatoriamente para cada árvore. Esse parâmetro ajuda a prevenir overfitting e pode melhorar a generalização do modelo;
- **Gamma**: controla a complexidade da árvore, especificando a redução mínima na função de perda necessária para fazer uma divisão. Valores maiores levam a árvores mais simples;
- **Subsample**: refere-se à fração de amostras a serem usadas para construir cada árvore. Reduzir esse valor pode levar a uma maior variação entre as árvores e ajudar a prevenir *overfitting*;
- **Min Child Weight**: define o peso mínimo da soma dos gradientes para dividir uma folha. Esse parâmetro impede a criação de nós muito pequenos, garantindo que cada divisão adicione valor significativo ao modelo.

O terceiro modelo empregado foi o *LightGBM*, outro algoritmo que também permite lidar com grandes conjuntos de dados e é conhecido por sua eficiência computacional e desempenho superior em relação a outros métodos de árvores de decisão. Para aplicação do *LightGBM*, os dados também foram divididos em conjuntos de treino e teste na proporção de 80/20, respectivamente. Após isso, também foi empregada a técnica de validação cruzada com 5 *k-fold* para garantir uma avaliação robusta dos modelos. Durante o processo de validação cruzada, os hiperparâmetros do modelo foram variados em cada *fold* e testados, permitindo uma busca abrangente pelas configurações mais adequadas. Essa abordagem é essencial para evitar o sobreajuste dos modelos aos dados de treinamento e garantir que eles generalizem bem para novos conjuntos de dados.

Os principais hiperparâmetros do *LightGBM* ajustados nos experimentos foram:

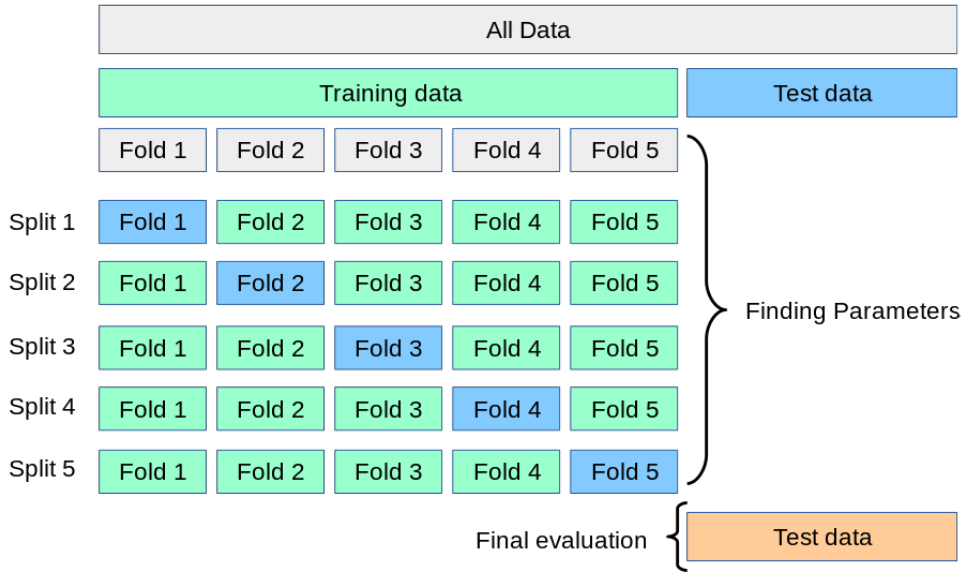
- **Num Leaves**: define o número máximo de folhas em cada árvore. Um número maior de folhas pode aumentar a capacidade do modelo de capturar relações complexas, mas também pode aumentar o risco de *overfitting*;

- **Learning Rate**: controla a taxa de aprendizado do modelo. Valores menores tornam o treinamento mais robusto, mas podem exigir um número maior de iterações para convergir;
- **Max Depth**: estabelece a profundidade máxima das árvores. Limitar a profundidade ajuda a evitar o *overfitting*, controlando a complexidade do modelo;
- **N Estimators**: especifica o número de árvores a serem construídas. Um maior número de árvores pode melhorar a precisão do modelo, mas também aumenta o tempo de treinamento;
- **Subsample**: refere-se à fração de amostras a serem usadas para construir cada árvore. Reduzir esse valor pode levar a uma maior variação entre as árvores e ajudar a prevenir *overfitting*;
- **Colsample Bytree**: representa a fração de colunas a serem amostradas aleatoriamente para cada árvore. Esse parâmetro ajuda a prevenir overfitting e pode melhorar a generalização do modelo;
- **Min Child Weight**: define o peso mínimo da soma dos gradientes para dividir uma folha. Esse parâmetro impede a criação de nós muito pequenos, garantindo que cada divisão adicione valor significativo ao modelo;
- **Reg Alpha**: controla a regularização L1 (Lasso) aplicada aos pesos. Aumentar esse valor pode simplificar o modelo, forçando alguns pesos a zero, o que ajuda a prevenir *overfitting*;
- **Reg Lambda**: controla a regularização L2 (*Ridge*) aplicada aos pesos. Aumentar esse valor ajuda a evitar *overfitting*, adicionando uma penalidade aos pesos elevados.

O ajuste desses hiperparâmetros, tanto para o *XGBoost* quanto para o *LightGBM*, é crucial e tem um impacto significativo no desempenho dos modelos. A importância da validação cruzada com *k-fold* reside na necessidade de testar faixas de valores para cada hiperparâmetro, a fim de identificar as melhores combinações, que equilibram a complexidade do modelo, com a sua capacidade de generalização. Em cada combinação de hiperparâmetros as métricas de desempenho são medidas para determinar a configuração mais eficiente e precisa, isso é essencial para obter previsões robustas e confiáveis.

Na Figura 10, está ilustrado o funcionamento do método de validação cruzada, mostrando a divisão dos dados em 5 *k-folds* para o treinamento dos modelos. Nesse processo, os 80% dos dados destinados ao treinamento são subdivididos em 5 partições iguais, cada uma contendo 16% dos dados originais. Essas partições são utilizadas iterativamente como conjuntos de treinamento e validação, enquanto os 20% restantes dos dados permanecem reservados para avaliação final do modelo.

Figura 10 – Diagrama de divisão de dados para validação cruzada com 5 *k-folds*.



Fonte: <https://scikit-learn.org/stable/modules/cross_validation.html>.

A utilização de replicações é fundamental para a coleta de amostras dos resultados dos experimentos, garantindo que as avaliações sejam estatisticamente robustas e confiáveis. Cada modelo foi avaliado com 10 replicações, que consistem na repetição do processo de treinamento e validação dos modelos, usando diferentes divisões dos dados. Esse procedimento é essencial para avaliar a estabilidade e a generalização dos modelos, proporcionando uma visão mais completa de seu comportamento em diferentes cenários.

Para a etapa de replicação no modelo de regressão linear, apenas é realizada a variação das porções de dados em que o modelo será aplicado a cada replicação. Já para os modelos *XGBoost* e *LightGBM* a cada replicação também são simuladas todas as combinações de hiperparâmetros pré-definidos. Assim, a necessidade de utilizar replicações está relacionada à realização conjunta de testes estatísticos que possam demonstrar diferenças significativas no desempenho dos modelos.

Para isso, utilizamos intervalos de confiança, que fornecem uma faixa de valores na qual o verdadeiro desempenho do modelo é esperado estar, com um certo nível de confiança.

Os intervalos de confiança para as médias das métricas de desempenho são calculados da seguinte forma:

- **Intervalo de Confiança Inferior**:

$$\text{IC Inferior} = \bar{x} - 1.96 \times \frac{\sigma}{\sqrt{10}}, \tag{3.4}$$

em que:

  - $\bar{x}$: é a Média das 10 replicações;

  - σ: é o Desvio Padrão das 10 replicações.

- **Intervalo de Confiança Superior**:

$$\text{IC Superior} = \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{10}}, \tag{3.5}$$

em que:

  - $\bar{x}$: é a Média das 10 replicações;
  - σ: é o Desvio Padrão das 10 replicações.

A aplicação de intervalos de confiança permite identificar se há uma sobreposição significativa nas faixas de desempenho dos diferentes modelos. Se os intervalos de confiança de dois modelos não se sobrepõem, isso indica uma diferença estatisticamente significativa entre seus desempenhos.

Outra ferramenta que foi implementada para a avaliação dos modelos foi a matriz de confusão. Em essência, a matriz de confusão é uma tabela que apresenta o número de previsões corretas e incorretas feitas por um modelo de classificação em relação a diferentes classes, facilitando a visualização dos erros específicos cometidos pelo modelo.

Neste trabalho, a matriz de confusão foi aplicada para avaliar o desempenho de três modelos preditivos - Regressão Linear, *XGBoost* e *LightGBM*. Para essa análise, foram definidas quatro faixas de valores de alvura para comparar os valores reais e preditos, conforme tabela 6.

Tabela 6 – Faixas de valores da Matriz de Confusão para comparação “real x predito”.

Faixas	**Intervalos de alvura para Matriz de Confusão**
1	< 69.0
2	69.0 a 71.0
3	71.0 a 73.0
4	> 73.0

Fonte: [o autor].

A utilização da matriz de confusão neste contexto oferece uma visão detalhada sobre como cada modelo distribui suas previsões entre as diferentes faixas de alvura. As matrizes foram elaboradas tanto em termos absolutos, apresentando o número total de ocorrências de dados em cada faixa, quanto em termos percentuais, demonstrando a proporção de dados em cada faixa. Esta abordagem dupla é metodologicamente importante porque permite uma avaliação não apenas da quantidade, mas também da distribuição relativa dos erros e acertos dos modelos, facilitando a comparação entre os diferentes métodos de predição.

Ao analisar as diagonais da matriz de confusão, é possível identificar a precisão com que os modelos classificam corretamente as faixas de alvura. Os elementos fora da diagonal fornecem informações sobre os tipos e frequências dos erros de classificação, indicando se há uma tendência específica dos modelos em superestimar ou subestimar a alvura em determinadas faixas. Essa análise detalhada é essencial para entender não apenas a acurácia global dos modelos, mas também para identificar áreas específicas de melhoria, como a necessidade de ajustar os modelos para lidar melhor com determinadas faixas de alvura.

3.7 Considerações Finais

Ao longo deste capítulo foram apresentadas as etapas e os procedimentos adotados para a construção e avaliação dos modelos preditivos de alvura no processo de branqueamento na indústria de celulose e papel. Inicialmente, discutimos sobre a importância da compreensão do problema e da definição dos objetivos do estudo, destacando a relevância da alvura como atributo chave nesse contexto industrial. Em seguida, são descritas a coleção de dados utilizada, abordando as principais variáveis de entrada e a variável de saída, bem como as características e peculiaridades do conjunto de dados.

A seleção de *features* foi realizada com base no conhecimento técnico das variáveis que impactam na alvura, neste processo das 13 variáveis carregadas inicialmente, apenas 8 seguiram para a etapa de correlação. Em seguida, a análise de correlação de *Pearson* foi essencial para identificar e remover variáveis altamente correlacionadas, a fim de evitar problemas de multicolinearidade e melhorar a eficiência do modelo. Assim, foi constatada a relação linear entre duas variáveis, resultando na exclusão da variável "teor de HexA na entrada do estágio DHT" devido à sua alta correlação com a variável "Kappa de entrada do Estágio DHT". Com essas etapas de pré-processamento foi possível reduzir o conjunto de variáveis de entrada para aquelas que apresentavam maior relevância e menor redundância.

A aplicação do PCA também foi considerada para a redução de dimensionalidade das variáveis de entrada, embora não tenha resultado em melhoria da performance dos modelos. Os experimentos foram definidos e configurados de maneira sistemática, seguindo a divisão dos dados entre treinamento e teste na proporção de 80/20, o que garantiu uma separação apropriada para a construção e validação dos modelos. A técnica de validação cruzada com 5 *k-folds* foi implementada para ajustar e otimizar os hiperparâmetros dos modelos *XGBoost* e *LightGBM*, permitindo uma busca detalhada pelas configurações mais adequadas.

Essa abordagem metodológica, com foco no pré-processamento dos dados, análise de correlação e seleção de *features*, bem como a definição de experimentos estruturados, foi essencial para garantir a robustez e a precisão dos modelos desenvolvidos. Essas etapas preparatórias criaram uma base sólida para a avaliação dos modelos preditivos de alvura, assegurando a confiabilidade dos resultados que serão apresentados.

4 Resultados

Este capítulo oferece uma análise detalhada dos resultados alcançados por este trabalho na aplicação de modelos de predição de alvura em um contexto industrial. Inicialmente, na Seção 4.1 são apresentados os resultados das métricas MAE, MSE e R^2 dos modelos de Regressão Linear, *XGBoost* e *LightGBM*. Na sequência, a Seção 4.2 descreve e interpreta os números obtidos e os padrões de comportamento dos modelos em diferentes faixas operacionais. A comparação com estudos anteriores é abordada na Seção 4.3 em que os resultados deste estudo são contextualizados com à literatura científica específica. Por fim, a Seção 4.4 explora como esses modelos podem ser aplicados para melhorar a eficiência operacional e guiar decisões estratégicas nas indústrias.

4.1 Desempenho dos Modelos de Predição

Para avaliar a eficácia dos modelos preditivos de alvura implementados: regressão linear, *XGBoost* e *LightGBM*, foram utilizadas as métricas: Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE) e o Coeficiente de Determinação (R^2). As justificativas para a escolha, bem como o funcionamento dessas métricas estão na seção de Metodologia deste trabalho.

Os dados de entrada para predição utilizados em todos os modelos consistiram em sete variáveis: Fluxo de Ácido, Produção do Branqueamento, Carga de Dióxido de Cloro, Teor de Shives, Kappa de Entrada do Estágio DHT, Alvura de Entrada do Estágio DHT e Temperatura de Entrada.

4.1.1 Regressão Linear

O desempenho do modelo de Regressão Linear pode ser avaliado com base nas métricas de avaliação de seu ajuste aos dados. Essas métricas estão apresentadas na Tabela 7, que demonstra na primeira coluna as 10 replicações executadas. Nas rês colunas seguintes ela apresenta os valores de MAE, MSE e R^2.

Ao final da tabela é realizado o cálculo da média, desvio padrão e dos intervalos de confiança de cada uma das métricas. Assim, o modelo apresentou um MAE médio de 0.911, um MSE médio de 1.324 e um R^2 médio de 0.599.

Esses resultados são considerados medianos e sugerem que o modelo de Regressão Linear proporcionou um ajuste razoável aos dados de entrada, capturando parte da relação entre as variáveis dependentes e independentes, embora não tenha capturado totalmente a complexidade dos dados.

Tabela 7 – Resultados para Regressão Linear das 10 replicações.

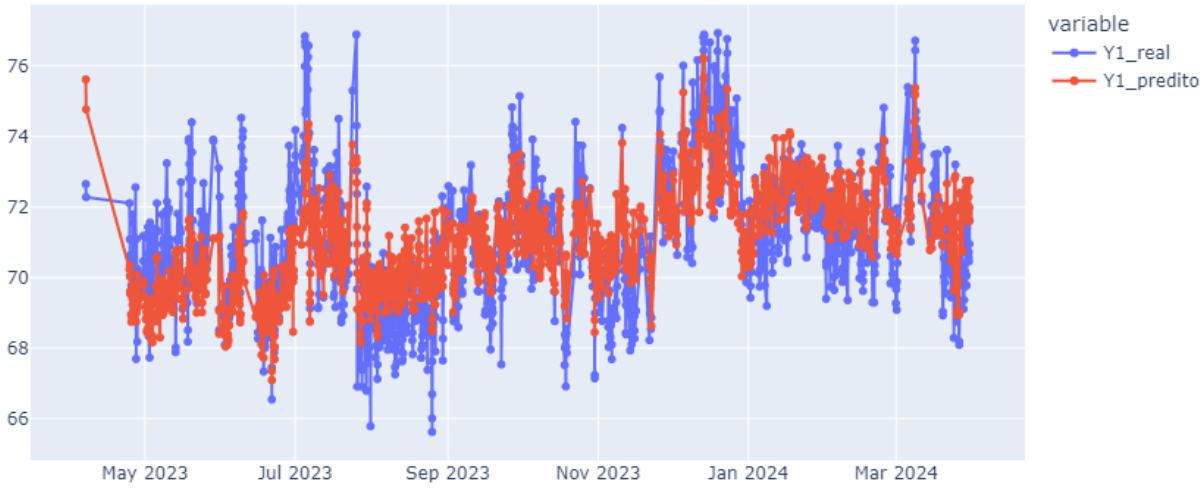
Regressão Linear			
Replicação	**MAE**	**MSE**	**R^2**
1	0.901	1.302	0.603
2	0.901	1.302	0.603
3	0.901	1.302	0.603
4	0.901	1.302	0.603
5	0.901	1.302	0.603
6	0.921	1.345	0.595
7	0.921	1.345	0.595
8	0.921	1.345	0.595
9	0.921	1.345	0.595
10	0.921	1.345	0.595
Média	**0.911**	**1.324**	**0.599**
Desvio Padrão	**0.011**	**0.023**	**0.004**
Intervalo de Confiança Inferior	**0.904**	**1.310**	**0.597**
Intervalo de Confiança Superior	**0.918**	**1.338**	**0.602**

Fonte: [o autor].

Várias hipóteses podem ser levantadas para explicar o resultado mediano do coeficiente de determinação (R^2) obtido pela Regressão Linear. Uma delas é que o relacionamento entre as variáveis independentes e dependentes pode não ser linear, o que limitaria a capacidade da Regressão Linear em modelar com precisão a relação entre elas. Além disso, a presença de *outliers* pode comprometer a capacidade do modelo de capturar adequadamente a estrutura dos dados, resultando em um ajuste menos preciso.

Na Figura 11 são exibidos os dados de "real *versus* predito" para o modelo de Regressão Linear. É possível notar que os dados reais atingem patamares superiores e inferiores mais extremos do que a Regressão Linear foi capaz de prever. A interpretação visual dos dados é importante e muito utilizada, embora as métricas apresentadas também tenham sido capazes de refletir o que se percebe visualmente.

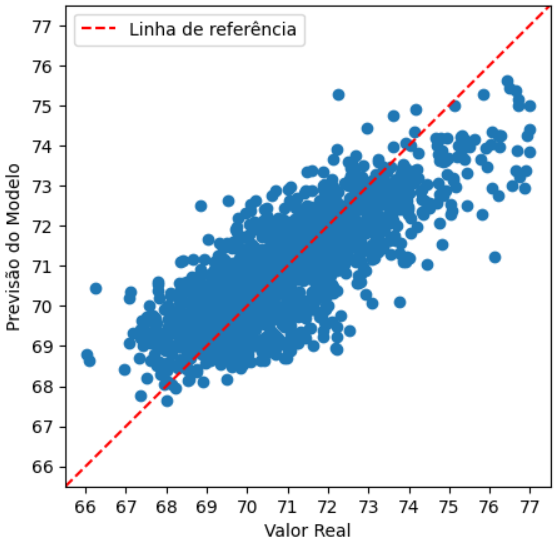
Figura 11 – Dados de "real *versus* predito" para o modelo de Regressão Linear.



Fonte: [O autor].

Na Figura 12 é apresentado o *Fitting* do modelo de Regressão Linear, a qual constitui outra forma de visualizar os dados de "real *versus* predito". Neste gráfico, é possível notar que os dados até seguem uma tendência de subida junto à linha de referência, entretanto ainda ficam muito dispersos, motivo pelo qual o R^2 apresentou apenas 0.599 de correlação.

Figura 12 – *Fitting* do modelo de Regressão Linear (Real *versus* Predito).

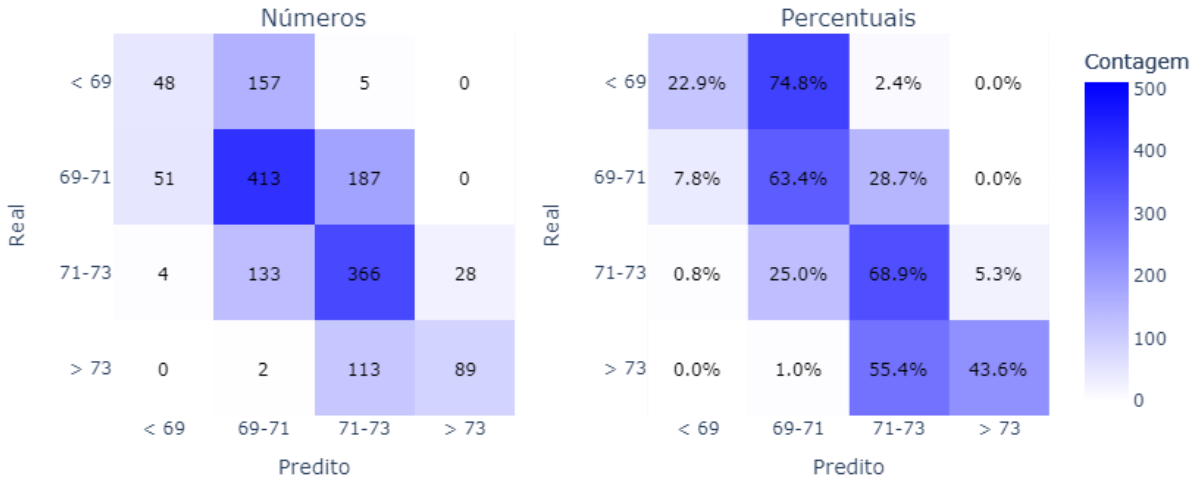


Fonte: [O autor].

Por fim, a Figura 13 apresenta as Matrizes de Confusão para a Regressão Linear, exibindo os dados tanto em termos totais quanto percentuais para cada faixa de alvura prevista. Essas faixas de alvura estão descritas na 6.

A matriz de confusão é uma ferramenta importante para quantificar a acurácia do modelo em cada faixa de alvura. Quanto mais dados estiverem na diagonal principal decrescente, maior será a assertividade do modelo. No entanto, devido à natureza linear do modelo de regressão, há uma tendência a cometer erros sistemáticos. Isso é evidenciado pelo percentual elevado de dados fora da diagonal principal decrescente, indicando que o modelo frequentemente falha em capturar a complexidade das relações entre as variáveis e a alvura prevista.

Figura 13 – Matriz de Confusão para Regressão Linear.



Fonte: [O autor].

4.1.2 XGBoost

Em relação ao modelo *XGBoost*, os resultados médios das métricas MAE, MSE e R^2, obtidos por meio de múltiplas replicatas do modelo com diferentes configurações de hiperparâmetros. Os resultados apresentam um desempenho geral bastante positivo, com um R^2 de aproximadamente 0.798. O MAE médio foi de aproximadamente 0.617 e o MSE médio foi de cerca de 0.670. Essas métricas de erro são relativamente baixas, o que é desejável em modelos de regressão.

Uma possível explicação para o desempenho positivo do modelo *XGBoost* pode ser atribuída à sua capacidade de lidar eficazmente com dados complexos e não-lineares. Esse algoritmo é conhecido por sua capacidade de capturar relações complexas entre variáveis de entrada e de saída, através da construção de árvores de decisão em série e da aplicação de técnicas avançadas de otimização.

Na Tabela 8 estão demonstrados os resultados de média e desvio padrão das 10 replicações simuladas. O intervalo de confiança das métricas fornece uma estimativa mais confiável do desempenho do modelo, levando em consideração a variação nos conjuntos de dados de treinamento e teste.

Tabela 8 – Melhores hiperparâmetros encontrados em cada replicação.

	XGBoost									
Rep.	**learning rate**	**max depth**	**n estimators**	**colsample bytree**	**gamma**	**subsample**	**min child weight**	**MAE**	**MSE**	**R²**
1	0.100	5.000	200.000	0.900	0.000	0.900	1.000	0.607	0.632	0.812
2	0.100	5.000	200.000	0.900	0.000	0.900	1.000	0.616	0.654	0.804
3	0.100	5.000	200.000	0.900	0.000	0.900	1.000	0.627	0.700	0.785
4	0.100	5.000	200.000	0.900	0.000	0.900	1.000	0.631	0.721	0.784
5	0.100	5.000	200.000	0.900	0.000	0.700	1.000	0.625	0.689	0.794
6	0.100	5.000	200.000	0.900	0.000	0.700	1.000	0.608	0.658	0.796
7	0.100	5.000	200.000	0.900	0.000	0.700	1.000	0.607	0.629	0.810
8	0.100	5.000	200.000	0.900	0.000	0.700	1.000	0.626	0.678	0.788
9	0.100	5.000	200.000	0.900	0.000	0.700	1.000	0.605	0.640	0.799
10	0.100	5.000	200.000	0.900	0.000	0.900	1.000	0.613	0.703	0.804
Média								**0.617**	**0.670**	**0.798**
Desvio Padrão								**0.010**	**0.032**	**0.010**
Intervalo de Confiança Inferior								**0.611**	**0.650**	**0.791**
Intervalo de Confiança Superior								**0.623**	**0.690**	**0.804**

Fonte: [o autor].

O procedimento de busca dos melhores hiperparâmetros do modelo *XGBoost* desempenha um papel fundamental na otimização da performance preditiva dos problemas. Essa abordagem sistemática visa explorar e identificar as configurações mais adequadas dos parâmetros do modelo, visando maximizar a acurácia das previsões.

Inicialmente, é essencial definir uma grade de hiperparâmetros a serem investigados. Essa grade abrange uma série de valores para cada parâmetro considerado, como a taxa de aprendizado, a profundidade máxima das árvores de decisão, o número de estimadores e outros. Tal diversidade de configurações permite uma ampla exploração do espaço de busca, possibilitando encontrar as combinações mais promissoras.

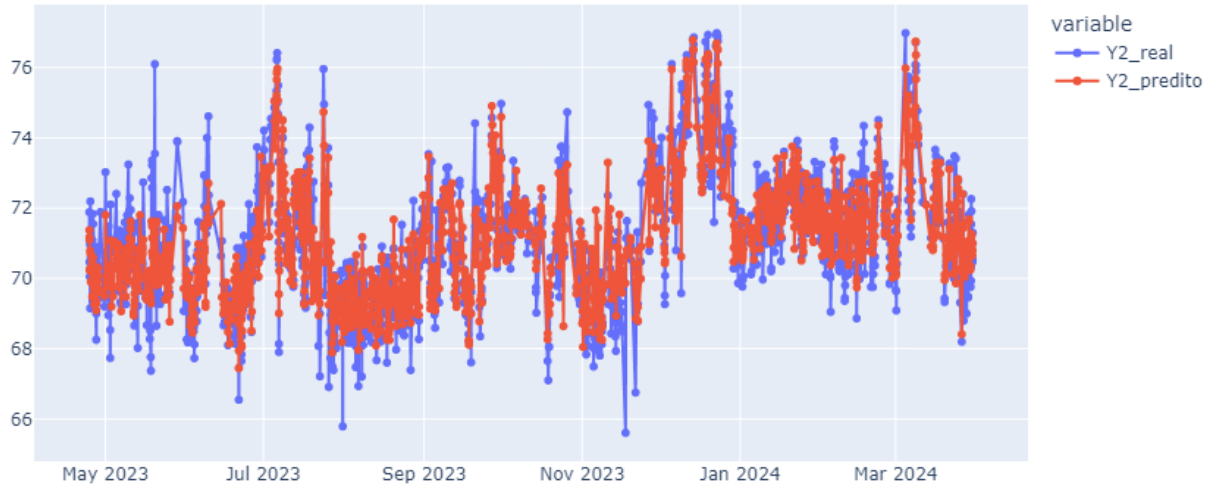
O processo de busca dos melhores hiperparâmetros é realizado por meio de técnicas de validação cruzada, empregadas pelo *GridSearchCV*. Essa abordagem divide o conjunto de dados em subconjuntos de treino e validação, permitindo avaliar o desempenho do modelo em diferentes combinações de hiperparâmetros. A seleção é guiada pela métrica de avaliação, neste caso o R^2, que quantifica a capacidade do modelo em explicar a variabilidade dos dados.

Além disso, o processo é repetido 10 vezes, com diferentes sementes aleatórias e combinações de hiperparâmetros, visando reduzir a dependência dos resultados em relação à aleatoriedade na divisão dos dados. Essa estratégia, conhecida como replicação, contribui para obter resultados mais robustos e confiáveis.

Uma vez identificados os melhores hiperparâmetros, o modelo *XGBoost* é treinado novamente utilizando todos os dados disponíveis. Isso assegura que o modelo final seja ajustado de forma ótima, utilizando a configuração mais promissora.

Na Figura 14 são apresentados os dados de "real *versus* predito" para o modelo *XGBoost*. É possível notar que os dados reais e preditos estão bem aderentes, indicando uma boa capacidade de predição do modelo. A interpretação visual dos dados é importante e muito utilizada, embora as métricas apresentadas também tenham sido capazes de refletir o bom desempenho do modelo.

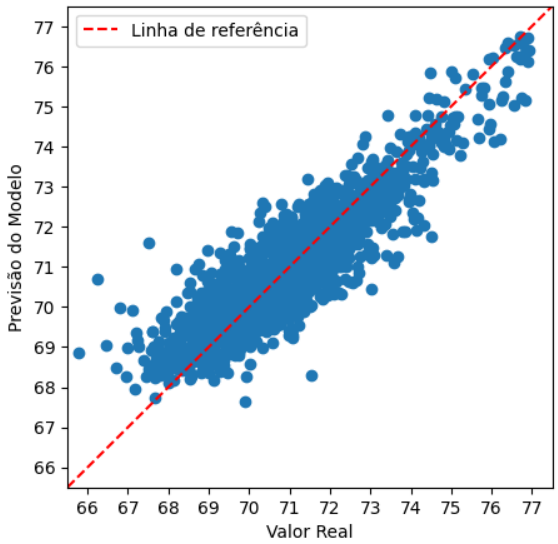
Figura 14 – Dados de "real *versus* predito" para o modelo *XGBoost*.



Fonte: [O autor].

Na Figura 15 é apresentado o *Fitting* do modelo *XGBoost*, a qual constitui outra forma de visualizar os dados de "real *versus* predito". Neste gráfico, é possível notar que os dados seguem uma adequada tendência de subida mais aderentes à linha de referência, motivo pelo qual o R^2 apresentou 0.798 de correlação.

Figura 15 – *Fitting* do modelo de *XGBoost* (Real *versus* Predito).

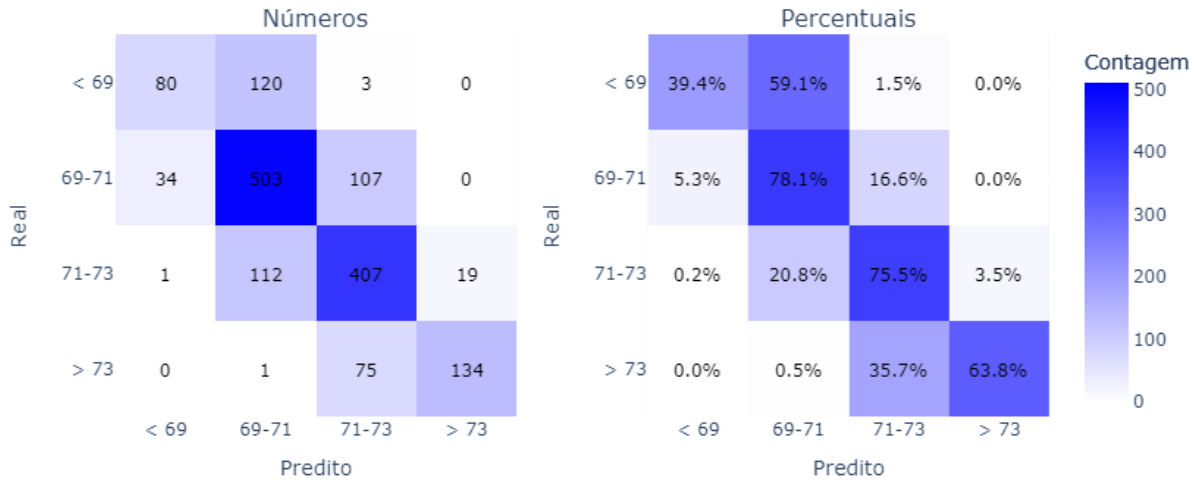


Fonte: [O autor].

Por fim, a Figura 16 apresenta as Matrizes de Confusão para o *XGBoost*, exibindo os dados tanto em termos totais quanto percentuais para cada faixa de alvura prevista. Essas faixas de alvura estão descritas na Tabela 6.

Neste caso, quanto mais dados estiverem na diagonal principal decrescente da matriz de confusão, maior será a assertividade do modelo. O *XGBoost* destaca-se por seu bom desempenho, conforme evidenciado pela concentração significativa de dados na diagonal principal decrescente. Isso indica que o modelo é eficaz em capturar a complexidade das relações entre as variáveis e a alvura prevista. A baixa quantidade de dados fora da diagonal principal decrescente ressalta a precisão do *XGBoost* em prever as faixas de alvura, superando os erros sistemáticos comumente associados a modelos lineares.

Figura 16 – Matriz de Confusão para *XGBoost*.



Fonte: [O autor].

4.1.3 LightGBM

Em relação ao modelo *LightGBM*, os resultados médios das métricas MAE, MSE e R^2, obtidos através de múltiplas replicatas do modelo com diferentes configurações de hiperparâmetros. Os resultados revelam um desempenho global bastante promissor, com um R^2 de aproximadamente 0.797. A média do MAE foi de cerca de 0.610 e do MSE foi de aproximadamente 0.661. Essas métricas indicam que o modelo apresenta uma boa capacidade de predição, com erros relativamente baixos, o que é altamente desejável em modelos de regressão.

Uma possível justificativa para o desempenho positivo do modelo *LightGBM* pode ser atribuída à sua eficácia em lidar com dados complexos e não-lineares. O algoritmo *LightGBM* é conhecido por sua capacidade de capturar relações complexas entre as variáveis de entrada e de saída, utilizando uma abordagem baseada em árvores de decisão em histogramas, o que lhe confere uma eficiência computacional superior em comparação com outros métodos baseados em árvores.

Na Tabela 9 estão demonstrados os resultados de média e desvio padrão das 10 replicações simuladas.Os resultados médios e o intervalo de confiança das métricas fornecem uma estimativa mais confiável do desempenho do modelo, levando em consideração a variação nos conjuntos de dados de treinamento e teste. Essas informações são essenciais para avaliar a generalização do modelo e sua capacidade de fazer previsões precisas em novos dados.

Tabela 9 – Melhores hiperparâmetros encontrados em cada replicação.

	LightGBM											
Rep.	**num leaves**	**learning rate**	**max depth**	**n estimators**	**subsample**	**colsample bytree**	**min child weight**	**reg alpha**	**reg lambda**	**MAE**	**MSE**	**R^2**
1	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.610	0.671	0.786
2	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.627	0.682	0.793
3	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.606	0.629	0.812
4	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.611	0.661	0.804
5	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.594	0.619	0.803
6	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.602	0.641	0.807
7	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.594	0.636	0.805
8	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.606	0.660	0.806
9	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.615	0.690	0.783
10	40.000	0.100	5.000	200.000	0.700	0.900	1.000	0.000	0.100	0.632	0.719	0.776
Média										**0.610**	**0.661**	**0.797**
Desvio Padrão										**0.013**	**0.031**	**0.012**
Intervalo de Confiança Inferior										**0.602**	**0.642**	**0.790**
Intervalo de Confiança Superior										**0.617**	**0.680**	**0.805**

Fonte: [o autor].

O procedimento de busca dos melhores hiperparâmetros do modelo *LightGBM* desempenha um papel crucial na otimização do desempenho preditivo do modelo. Essa abordagem sistemática visa explorar e identificar as configurações mais adequadas dos parâmetros do modelo, visando maximizar a precisão das previsões.

Para realizar essa busca, é essencial definir uma grade de hiperparâmetros a serem investigados, abrangendo uma série de valores para cada parâmetro considerado, como a taxa de aprendizado, a profundidade máxima das árvores de decisão, o número de estimadores, entre outros. Essa variedade de configurações permite uma exploração abrangente do espaço de busca, facilitando a identificação das combinações mais promissoras.

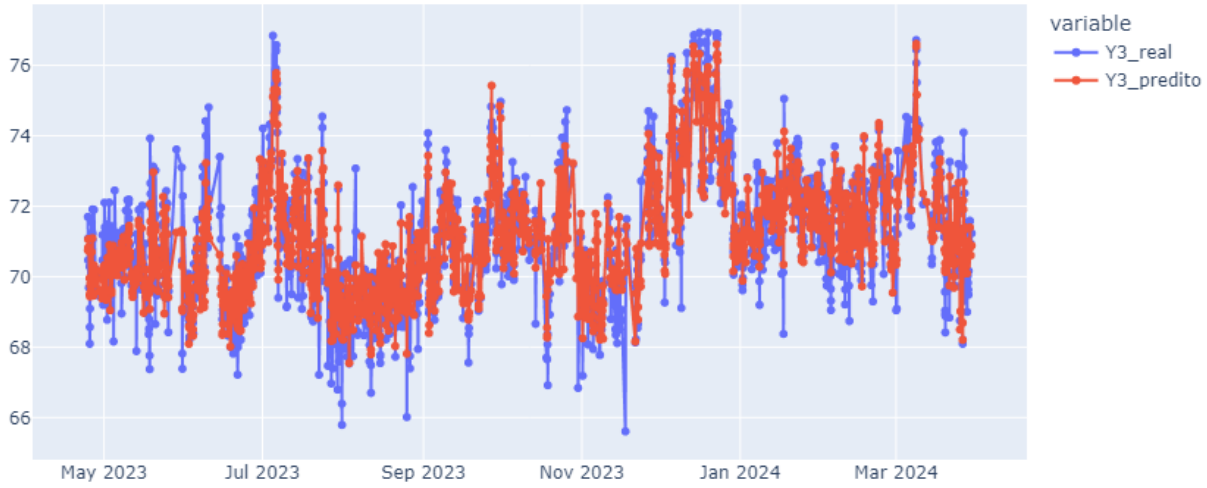
O processo de busca dos melhores hiperparâmetros é conduzido utilizando técnicas de validação cruzada, empregadas pelo *GridSearchCV*. Essa abordagem divide o conjunto de dados em subconjuntos de treino e validação, permitindo avaliar o desempenho do modelo em diferentes combinações de hiperparâmetros. A seleção é guiada pela métrica de avaliação, neste caso o R^2, que quantifica a capacidade do modelo em explicar a variabilidade dos dados.

Além disso, o processo é repetido 10 vezes, com diferentes sementes aleatórias e combinações de hiperparâmetros, visando reduzir a dependência dos resultados em relação à aleatoriedade na divisão dos dados. Essa estratégia, conhecida como replicação, contribui para obter resultados mais robustos e confiáveis.

Uma vez identificados os melhores hiperparâmetros, o modelo *LightGBM* é treinado novamente utilizando todos os dados disponíveis. Isso garante que o modelo final seja ajustado de forma ótima, utilizando a configuração mais promissora.

Na Figura 17 observa-se a comparação entre os valores reais e os valores previstos pelo modelo *LightGBM*. A proximidade entre os pontos reais e preditos ilustra a alta precisão do modelo, demonstrando sua eficácia em capturar a variabilidade dos dados. A análise visual é fundamental para compreender a capacidade do modelo em prever com exatidão, complementando as métricas quantitativas que também confirmam o bom desempenho do *LightGBM*. A figura destaca a aderência dos dados, evidenciando a robustez do modelo em fornecer previsões confiáveis.

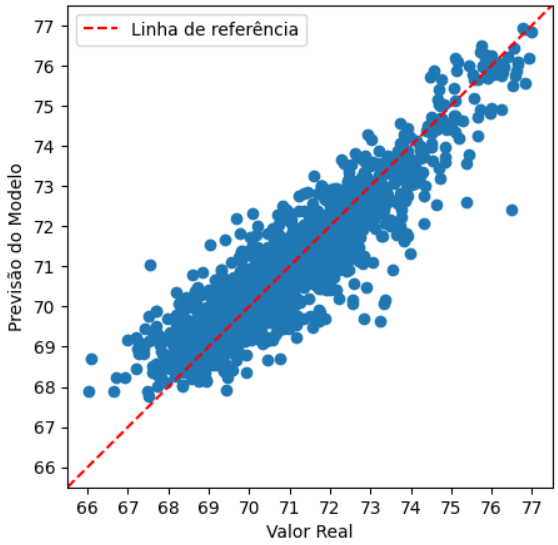
Figura 17 – Dados de "real *versus* predito" para o modelo *LightGBM*.



Fonte: [O autor].

Na Figura 18 é possível visualizar o *Fitting* do modelo *LightGBM*, que exibe a relação entre os valores reais e preditos. Neste gráfico, a proximidade dos pontos à linha de referência indica uma boa capacidade preditiva do modelo. A aderência dos dados à linha de referência reflete a eficácia do *LightGBM* em capturar as tendências subjacentes, resultando em um R^2 de 0.797. Esse valor de R^2 destaca a boa correlação e precisão do modelo, evidenciando seu desempenho superior em comparação com modelos menos complexos.

Figura 18 – *Fitting* do modelo de *LightGBM* (Real *versus* Predito).

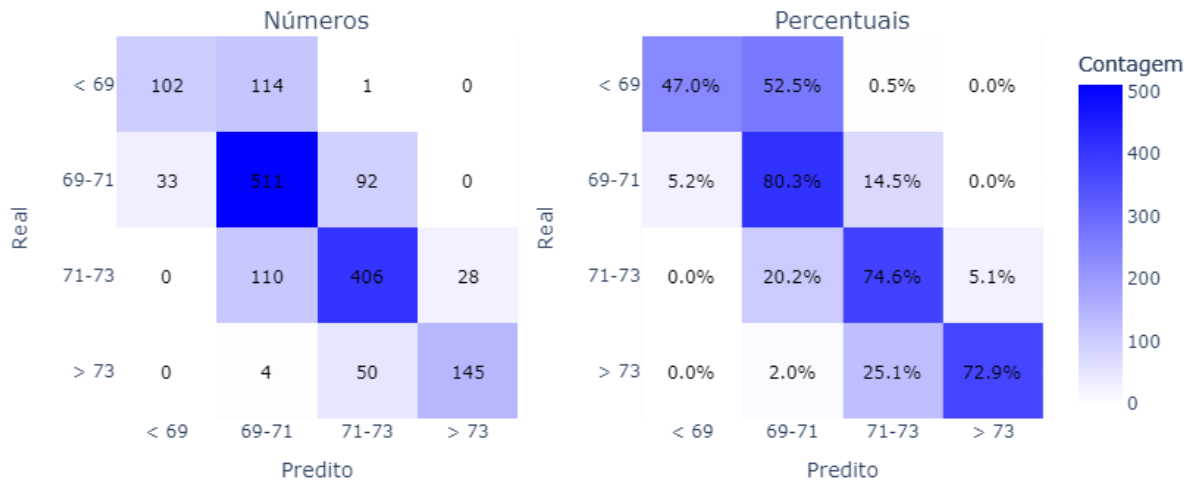


Fonte: [O autor].

Por fim, a Figura 19 são apresentadas as Matrizes de Confusão para o *LightGBM*, exibindo os resultados tanto em termos totais quanto percentuais para cada faixa de alvura prevista, conforme descrito na Tabela 6.

A interpretação da matriz de confusão indica que quanto mais dados estiverem na diagonal principal decrescente, maior será a precisão do modelo. O *LightGBM* demonstra um excelente desempenho, como evidenciado pela alta concentração de dados nessa diagonal. Isso sugere que o modelo é eficaz em capturar as nuances das relações entre as variáveis e a alvura prevista. A baixa quantidade de dados fora da diagonal principal decrescente ressalta a capacidade do *LightGBM* em prever com precisão as faixas de alvura, superando as limitações frequentemente encontradas em modelos lineares.

Figura 19 – Matriz de Confusão para *LightGBM*.



Fonte: [O autor].

4.2 Análise dos Resultados

Com base nos resultados apresentados na Tabela 10, podemos iniciar a discussão sobre os modelos Regressão Linear, *XGBoost* e *LightGBM*, destacando suas vantagens e limitações, bem como a identificação de padrões ou tendências nos dados preditos e observados.

Após treinar e testar os modelos com os dados do processo, os resultados mostraram que o modelo de regressão linear apresentou uma performance moderada. Por outro lado, o *XGBoost* e o *LightGBM* demonstraram um desempenho muito superior, sugerindo que os modelos baseados em árvores de decisão são mais adequados para realizar esta predição a alvura no estágio.

Os resultados completos comparativos das métricas de teste em cada um dos modelos estão apresentados na Tabela 10. A Regressão Linear apresentou um MAE, que mede o erro médio absoluto entre os valores previstos e reais, de 0.911. O MSE foi de 1.324, indicando o erro quadrático médio, que penaliza mais os grandes erros. O R^2 foi de 0.599, representando a proporção da variância na variável dependente que é previsível a partir da variável independente.

O *XGBoost* demonstrou uma significativa superioridade nas métricas em comparação com a Regressão Linear. O MAE do *XGBoost* foi 32.27% menor do que o da Regressão Linear, o MSE foi 49.40% menor, e o R^2 foi 24.94% maior. Esses resultados evidenciam que o *XGBoost* conseguiu reduzir consideravelmente os erros de previsão e aumentar a explicabilidade em comparação com a abordagem linear da Regressão Linear.

Similarmente, o *LightGBM* também se destacou em relação à Regressão Linear. O MAE do *LightGBM* foi 33.04% menor, o MSE foi 50.08% menor, e o R^2 foi 24.84% maior. Esses resultados ressaltam a capacidade do *LightGBM* em prever com maior precisão e explicar uma maior proporção da variância nos dados de alvura, superando as limitações da abordagem linear da Regressão Linear.

Tabela 10 – Resultado das métricas dos Modelos.

Modelo	**MAE**	**MSE**	**R^2**
Regressão Linear	0.911	1.324	0.599
XGBoost	0.617	0.670	0.798
LightGBM	0.610	0.661	0.797

Fonte: [o autor].

Conforme descrito, O modelo de Regressão Linear apresentou um desempenho relativamente inferior em comparação com os modelos *XGBoost* e *LightGBM*, em todas as métricas avaliadas. No entanto, uma vantagem da Regressão Linear é sua simplicidade e interpretabilidade, o que pode ser útil para entender a relação entre as variáveis de forma mais intuitiva.

Por outro lado, os modelos de *XGBoost* e *LightGBM* mostraram um desempenho significativamente melhor em todas as métricas avaliadas. No entanto, uma desvantagem desses modelos é sua maior complexidade e dificuldade de interpretação em comparação com a regressão linear.

Ao analisar os padrões nos dados preditos e observados, podemos observar que tanto o *XGBoost* quanto o *LightGBM* foram capazes de capturar padrões mais precisos nos dados, como evidenciado pelos valores mais baixos de MAE e MSE e pelos valores mais altos de R^2. Isso sugere que esses modelos foram mais eficazes em prever a variável alvo com base nas variáveis preditoras.

Em resumo, enquanto a Regressão Linear oferece simplicidade e interpretabilidade, os modelos de *ensemble* como *XGBoost* e *LightGBM* superam em termos de desempenho preditivo. A escolha entre esses modelos depende das necessidades específicas do problema, considerando a precisão preditiva de cada um.

A análise comparativa de modelos de regressão é essencial para a compreensão do desempenho de diferentes algoritmos na modelagem de fenômenos complexos. Os resultados apresentados na Tabela 11 mostram os intervalos de confiança para cada métrica, tanto para a Regressão Linear quanto para os modelos baseados em árvores de decisão, *XGBoost* e *LightGBM*.

Os intervalos de confiança para as métricas (MAE, MSE e R²) dos modelos, foram calculados para fornecer uma estimativa da precisão dos resultados. Eles foram determinados com um nível de confiança de 95%, o que implica uma significância estatística de 5%. Esses intervalos foram calculados utilizando a média e o desvio padrão das dez replicações, conforme as Equações 3.4 e 3.5.

Tabela 11 – Resultado dos Intervalos de Confiança métricas dos Modelos.

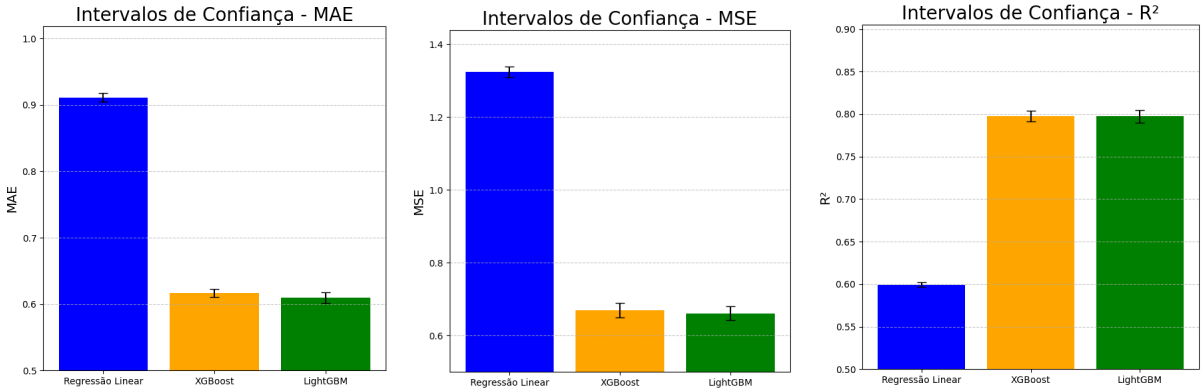
Modelos	**Intervalos**	**MAE**	**MSE**	**R²**
Regressão Linear	Intervalo de Confiança Inferior	0.904	1.310	0.597
	Intervalo de Confiança Superior	0.918	1.338	0.602
XGBoost	Intervalo de Confiança Inferior	0.611	0.650	0.791
	Intervalo de Confiança Superior	0.623	0.690	0.804
LightGBM	Intervalo de Confiança Inferior	0.602	0.642	0.790
	Intervalo de Confiança Superior	0.617	0.680	0.805

Fonte: [o autor].

Os resultados da Tabela 11 estão melhor apresentados de forma gráfica nas Figuras 20a, 20b e 20c. Notavelmente, a Regressão Linear exibe intervalos de confiança mais amplos em comparação com os modelos baseados em árvores de decisão, *XGBoost* e *LightGBM*. Isso demonstra que a Regressão Linear possui uma estimativa menos precisa das métricas de desempenho em comparação com os modelos de árvores de decisão.

Ao comparar especificamente os modelos *XGBoost* e *LightGBM*, observamos que seus intervalos de confiança para todas as métricas se sobrepõem, indicando que não há diferenças estatisticamente significativas entre esses dois modelos em termos de desempenho. Isso sugere que ambos os modelos *XGBoost* e *LightGBM* podem ser igualmente adequados para o problema em questão, com base nas métricas avaliadas.

Figura 20 – Comparação dos Intervalos de Confiança entre os modelos para diferentes métricas.



(a) Comparação dos Intervalos de Confiança entre os modelos para a métrica MAE.

(b) Comparação dos Intervalos de Confiança entre os modelos para a métrica MSE.

(c) Comparação dos Intervalos de Confiança entre os modelos para a métrica R^2.

Fonte: [O autor].

4.3 Comparação com a Literatura

A análise comparativa dos modelos preditivos desenvolvidos neste estudo se alinha com descobertas anteriores na indústria de celulose e papel. Estudos prévios indicam que modelos baseados em árvores de decisão, como o *XGBoost* e o *LightGBM*, tendem a superar modelos lineares em termos de precisão preditiva em cenários complexos. Por exemplo, Chen e Guestrin (2016) mostraram que o *XGBoost* é particularmente eficaz para lidar com grandes volumes de dados e identificar relações não lineares.

Este resultado é corroborado por Chen *et al.* (2019), que destacaram a robustez do *XGBoost* em contextos industriais similares, sugerindo que este modelo pode capturar padrões complexos e proporcionar previsões precisas.

Os resultados apresentados em nosso estudo, onde o *XGBoost* e o *LightGBM* superaram significativamente a Regressão Linear, confirmam as observações de outros trabalhos na literatura. O modelo *LightGBM*, desenvolvido por Ke *et al.* (2017), demonstrou ser altamente eficiente em termos de tempo de treinamento e capacidade de generalização, características que também foram observadas em nossa análise.

A aplicação da Análise de Componentes Principais (PCA) foi explorada para a redução de dimensionalidade e visualização dos dados. No entanto, os resultados mostraram que a inclusão das componentes principais (PC1, PC2 e PC3) nos modelos não resultou em um desempenho satisfatório. Optamos, portanto, por utilizar as sete variáveis originais. Esta decisão está em consonância com os argumentos de Jolliffe (2002), que indicam que a eficácia do PCA depende fortemente da natureza dos dados e da quantidade de variáveis originais. Em nosso caso, o número de variáveis não foi suficientemente grande para justificar a redução dimensional significativa.

Essas observações são importantes ao considerar a aplicabilidade prática dos modelos preditivos na indústria de celulose e papel. A Regressão Linear, embora mais simples e interpretável, não capturou a complexidade dos dados do processo. Em contrapartida, os modelos *XGBoost* e *LightGBM* mostraram-se mais adequados para prever a alvura no estágio, devido à sua capacidade de lidar com interações não lineares e variabilidade nos dados. Essa conclusão é consistente com as expectativas baseadas na literatura, indicando que os modelos baseados em árvores de decisão continuam a ser uma escolha robusta para problemas preditivos complexos na indústria.

4.4 Implicações Práticas

Os modelos *XGBoost* e *LightGBM* demonstraram desempenhos equivalentes nos resultados obtidos. No entanto, optou-se pela implementação do modelo preditivo *XGBoost* no *PI System* da unidade industrial, devido ao R^2 ligeiramente superior de 0.798 em comparação com 0.797.

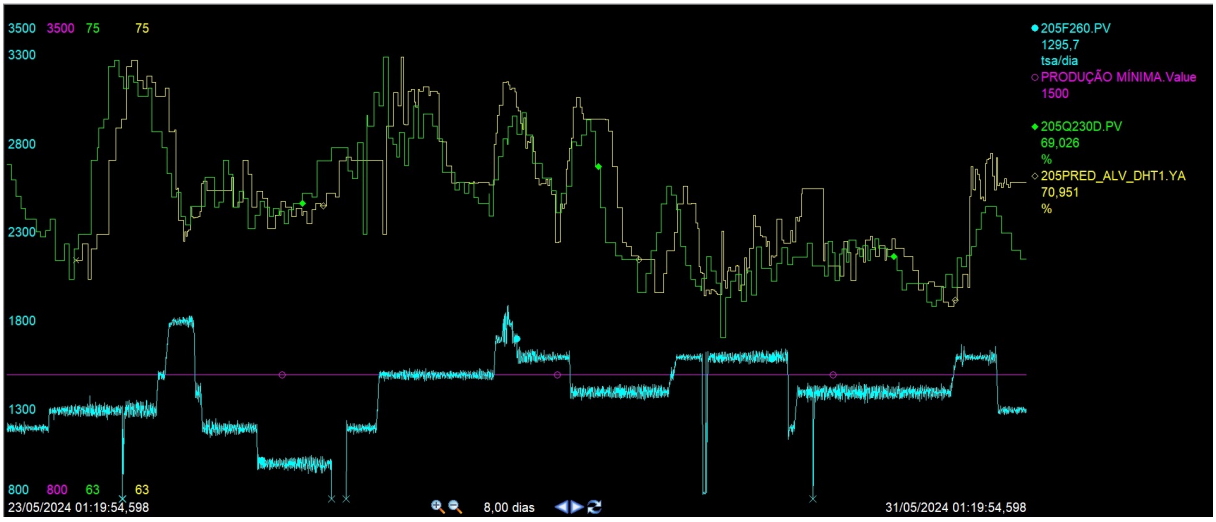
A implementação do *XGBoost* ao *PI System* possibilita a geração de previsões de forma *online*. Assim, o sistema captura continuamente os dados de entrada e fornece as previsões de saída, que são comparadas com os dados reais medidos pelos instrumentos, como demonstrado na Figura 21.

O PI System, desenvolvido pela OSIsoft, agora parte da AVEVA, é uma plataforma robusta de gerenciamento de dados em tempo real, amplamente utilizada em unidades industriais para otimizar operações e melhorar a tomada de decisões. Este sistema coleta, armazena e gerencia grandes volumes de dados provenientes de diversas fontes, incluindo sensores de campo, controladores de processos como Programmable Logic Controller (PLC), Sistema Digital de Controle Distribuído (SDCD) e outros sistemas de TI.

O *PI Data Archive*, núcleo do PI System, é um banco projetado para armazenar milhões de pontos de dados com alta eficiência e fidelidade temporal, funcionando como um historiador que registra continuamente dados de processos para análises retrospectivas. Ferramentas como *PI Vision* e *PI DataLink*, amplamente utilizadas na fábrica, proporcionam a visualização interativa e integração com *Microsoft Excel*, possibilitando a criação de *dashboards* e relatórios personalizados para monitoramento em tempo real e análise de tendências.

Em nossas aplicações práticas o PI System os coleta dados das variáveis de processo em tempo real, esses dados são armazenados para análises históricas e utilizados para desenvolver os modelos preditivos deste trabalho. Esse sistema se destaca pela alta capacidade de integrar-se com plataformas de *machine learning* e *big data* para análises avançadas e desenvolvimento de modelos preditivos.

Figura 21 – Comparativo da alvura real medida (pena verde) ao lado do valor predito pelo modelo (pena amarela).



Fonte: PI System.

Neste gráfico é possível visualizar a predição de alvura do modelo (pena amarela) e compará-lo com o valor real medido pelo instrumento (pena verde). Na parte inferior do gráfico também são exibidas a produção do branqueamento (pena azul), junto do limite de produção mínima aceitável pelo modelo de 1500 tSA/d (pena roxa). Assim, é possível notar que nos momentos em que a produção do branqueamento (pena azul) opera acima da produção miníma (pena roxa), o modelo é capaz de antecipar a alvura em até 3 horas, que é exatamente o tempo de retenção da torre DHT.

Devido algumas instabilidades na unidade durante o período observado, a produção do branqueamento A variou muito e operou abaixo de 1500 tSA/d em muitos momentos. Isso nos permite confirmar que o modelo não é capaz de antecipar a alvura de saída nos momentos em que as variáveis de entrada saem dos patamares filtrados, conforme estabelecidos na Tabela 2.

Neste primeiro momento foi priorizada a implementação do controle em cenários de produção mais estável no branqueamento A. Em trabalhos futuros o limite inferior do filtro de produção será reduzido e o modelo retreinado. Isso poderá impactar um pouco na assertividade do controle, pois irá inserir ao modelo dados menos sadios do processo, com a produção reduzida.

A implementação desse modelo no PI System representou um avanço significativo e entre os próximos passos está a integração ao controle de químicos do branqueamento A. Algumas implicações práticas são esperadas:

- **Otimização de Processos**: A capacidade de prever a alvura permite ajustes no processo de branqueamento em tempo real, garantindo especificações de qualidade de forma consistente. Isso resulta em um processo mais eficiente e menor variabilidade na qualidade do produto;
- **Tomada de Decisão Baseada em Dados**: A implementação do modelo preditivo fornece aos operadores e engenheiros de processo uma ferramenta para tomar decisões embasadas em dados e previsões confiáveis. Durante momentos de instabilidade, quando é necessário o controle manual, os operadores terão maior facilidade em ajustar a carga de químicos, observando a resposta do preditor;
- **Redução de Custo Variável**: Antecipar a alvura permitirá reduzir o consumo de químicos de branqueamento em momentos que a qualidade da polpa de entrada melhorar;
- **Melhoria da Qualidade do Produto**: Com previsões precisas, é possível manter a alvura do produto final dentro dos limites desejados, aumentando a carga de químicos apenas quando a qualidade da polpa de entrada assim demandar.

Em resumo, a aplicação de modelos preditivos como o *XGBoost* tem um potencial de impactar significativamente na otimização dos processos e na melhoria da qualidade do produto final. Ao antecipar os resultados de alvura com base em dados de produção em tempo real, a empresa pode melhorar a eficiência operacional, reduzir custos e aumentar a satisfação do cliente. Este estudo demonstra o potencial das tecnologias de análise de dados avançadas para transformar operações industriais e proporcionar vantagens competitivas.

5 Conclusões e trabalhos futuros

Para alcançar os objetivos específicos e o objetivo geral do trabalho, foram desenvolvidas várias etapas metodológicas detalhadas. Primeiramente, realizou-se a coleta de dados operacionais da unidade industrial. Esses dados foram essenciais para a construção de um banco de dados robusto que serviu como base para todas as análises subsequentes. Em seguida, foi realizada uma análise exploratória dos dados para caracterizar os dados da coleção de dados, identificar *outliers* e analisar correlações entre as variáveis. Isso permitiu uma compreensão inicial do comportamento do processo e ajudou a direcionar o desenvolvimento dos modelos preditivos.

A terceira etapa consistiu na seleção e implementação de três modelos de aprendizado de máquina: Regressão Linear, *XGBoost* e *LightGBM*. Cada modelo foi treinado e validado utilizando técnicas adequadas de validação cruzada e otimização de hiperparâmetros. Os resultados obtidos, incluindo métricas como MAE, MSE e R^2, foram cuidadosamente analisados para avaliar o desempenho de cada modelo na previsão da variável alvo, a alvura do produto final.

Na sequência, os modelos foram comparados entre si e com estudos anteriores da literatura científica, destacando as vantagens e limitações de cada abordagem. Essa comparação permitiu validar a robustez dos modelos desenvolvidos e contextualizar os resultados dentro do conhecimento existente. Por fim, o modelo *XGBoost*, de melhor desempenho pelas métricas, foi integrado ao sistema de gestão de dados em tempo real da unidade industrial (PI System), permitindo a implementação das previsões geradas em tempo real no processo de produção. Essa aplicação prática demonstrou como os modelos podem ser utilizados para melhorar a eficiência operacional e auxiliar na tomada de decisões estratégicas.

5.1 Resultados e Contribuições

O principal resultado do trabalho foi demonstrar a viabilidade de modelos preditivos de alvura em estágios de branqueamento, além de ter encontrado as principais variáveis de entrada úteis para essa predição. Ao aplicar três modelos de aprendizado de máquina: Regressão Linear, *XGBoost* e *LightGBM*, o trabalho também sugere bons modelos que podem ser utilizados nesse trabalho. Os resultados quantitativos demonstram a superioridade dos modelos baseados em árvores de decisão em relação ao modelo de Regressão Linear.

A Regressão Linear apresentou um MAE de 0.911, MSE de 1.324 e R^2 de 0.599, indicando uma performance mediana. Em contraste, o *XGBoost* demonstrou ser 32.27% melhor em MAE, 49.40% melhor em MSE e 24.94% melhor em R^2 em comparação com a Regressão Linear. Similarmente, os resultados mostraram que o *LightGBM* também se destacou ao ser 33.04% mais preciso em MAE, 50.08% melhor em MSE e 24.84% superior em R^2, em relação à Regressão Linear.

Os principais *insights* obtidos a partir dos resultados incluem a constatação de que a relação entre as variáveis independentes e a variável dependente (alvura) não é perfeitamente linear, o que limita a eficácia da Regressão Linear. Além disso, a aplicação de técnicas de pré-processamento, como a remoção de *outliers* e normalização dos dados com a transformação logarítmica, foi fundamental para melhorar a qualidade das previsões. No entanto, a utilização do PCA não resultou em melhorias significativas, sugerindo que a complexidade e correlação das sete variáveis originais poderiam ser melhor gerenciadas diretamente pelos modelos preditivos sem a necessidade de redução dimensional.

As contribuições deste trabalho são interessantes tanto para a indústria de celulose quanto para a literatura acadêmica. Na indústria, a implementação do modelo *XGBoost* no *PI System* da unidade industrial demonstra a viabilidade prática de integrar modelos preditivos avançados em processos de produção em tempo real, permitindo previsões online da alvura que são comparadas continuamente com os dados reais medidos. Esta integração pode resultar em otimização do processo, redução de custo variável e melhorias na qualidade do produto final.

Para a literatura, este estudo reforça a eficácia dos modelos baseados em árvores de decisão, como o *XGBoost* e o *LightGBM*, em cenários industriais complexos. A análise comparativa dos modelos mostrou que essas técnicas de aprendizado de máquina são capazes de capturar relações não lineares e complexas entre variáveis, oferecendo uma alternativa robusta à Regressão Linear tradicional. Além disso, os resultados confirmam a importância do pré-processamento dos dados e da seleção adequada de variáveis para a melhoria do desempenho dos modelos preditivos.

Em resumo, o estudo contribui com evidências práticas e teóricas de que a aplicação de técnicas avançadas de aprendizado de máquina pode trazer melhorias significativas na previsão de parâmetros críticos em processos industriais, promovendo uma maior eficiência e qualidade nas operações.

5.2 Limitações do Trabalho

Uma das principais limitações deste trabalho está relacionada aos filtros aplicados nas variáveis durante o pré-processamento dos dados. Esses filtros, destinados a remover inconsistências e *outliers*, estabelecem limites para as variáveis de entrada do modelo preditivo. Quando os valores das variáveis excedem esses limites, o modelo perde a capacidade de antecipar corretamente a alvura, resultando em um dado atrasado em relação à análise real. Esse problema se manifesta especialmente quando há variação de produção do branqueamento A, que além de potencialmente impactar todas as demais variáveis, ela também se configura como um filtro.

Para superar essa limitação, é necessário adotar uma abordagem progressiva de refinamento dos limites estabelecidos para as variáveis, sendo necessário a redução gradual desses limites, aliada ao re-treinamento contínuo do modelo. Isso permitirá que o modelo se adapte melhor às variações reais dos dados do processo e poderá ser feito através de um ciclo iterativo de reavaliação e ajuste. Dessa forma, o modelo pode ser calibrado de maneira mais precisa, reduzindo a probabilidade de exclusão de dados relevantes e avaliando a capacidade de lidar com dados menos sadios.

Além disso, a exploração de técnicas alternativas de pré-processamento pode oferecer soluções complementares para essa limitação. Métodos como a análise robusta de *outliers*, que identifica e trata os *outliers* de maneira diferenciada, sem necessariamente removê-los, poderiam ser investigados. Outra alternativa é a utilização de transformações de dados mais sofisticadas que normalizam as variáveis sem perder informações essenciais, como transformações baseadas em técnicas de aprendizagem profunda que preservam a estrutura e a distribuição original dos dados.

Por fim, futuros trabalhos podem se beneficiar da incorporação de variáveis adicionais que capturam aspectos do processo que não foram considerados neste estudo. A inclusão de novos dados pode enriquecer o modelo, proporcionando uma visão mais abrangente e detalhada do processo de branqueamento de celulose. Também é recomendável explorar modelos preditivos mais avançados, como redes neurais profundas e métodos de ensemble mais complexos, que podem oferecer melhor desempenho na presença de dados variáveis e complexos.

5.3 Trabalhos Futuros

Na continuação deste trabalho, uma linha de investigação promissora envolve a redução gradual dos limites estabelecidos no pré-processamento de dados. Essa abordagem provou ser eficaz para garantir a qualidade dos dados de entrada e, consequentemente, das previsões do modelo. No entanto, uma redução mais controlada e gradual desses limites pode ser explorada para avaliar o impacto direto na performance dos modelos. Esse ajuste fino permitirá identificar o ponto de equilíbrio entre a remoção de ruídos e a preservação de informações relevantes, possibilitando um aumento na precisão do modelo preditivo.

Além disso, a retreinamento periódico do modelo preditivo é essencial. O ambiente industrial é dinâmico, e as condições operacionais podem variar ao longo do tempo. Portanto, é crucial que o modelo seja continuamente atualizado com novos dados para manter sua relevância e acurácia. Esse processo de retreinamento pode ser automatizado, utilizando um sistema de monitoramento contínuo que identifica quando a performance do modelo começa a degradar, acionando assim um novo ciclo de treinamento com os dados mais recentes.

A integração do preditor com o sistema de controle de químicos do processo de branqueamento de celulose também é uma área de interesse significativa. Atualmente, o modelo preditivo fornece estimativas que podem ser utilizadas para ajustes manuais no processo. No futuro, o objetivo é implementar um sistema de *feedback*, onde as previsões de alvura influenciam diretamente a dosagem de dióxido de cloro e outros reagentes, pode otimizar o consumo de químicos e melhorar a qualidade do produto final. Essa integração requer a harmonização dos dados preditivos com os algoritmos de controle de processos existentes, criando um sistema de controle adaptativo que responda em tempo real às previsões do modelo.

Outra vertente importante para trabalhos futuros é a exploração de novas técnicas de aprendizado de máquina e modelagem preditiva. Embora o *XGBoost* e o *LightGBM* tenham mostrado excelente desempenho, outras técnicas avançadas de *machine learning*, como redes neurais profundas (*deep learning*) mais complexos poderiam ser investigados. Essas técnicas poderiam potencialmente capturar as não linearidades e interações complexas presentes nos dados de processo.

Outra abordagem para seleção de variáveis que será aplicada futuramente é a *Backward Elimination*, com o objetivo de identificar as mais significativas para a predição da alvura. Neste método as variáveis são removidas, uma por vez e de forma iterativa, e o desempenho do modelo é recalculado. Esse processo continua até que todas as variáveis restantes no modelo atendam a um critério de significância pré-definido. A aplicação do *Backward Elimination* no contexto da predição de alvura pode simplificar o modelo, eliminando variáveis redundantes ou irrelevantes, resultando em modelos mais robustos e interpretáveis. Esse método é particularmente útil para evitar o *overfitting*, garantindo que o modelo final inclua apenas as variáveis que realmente contribuem para a precisão das previsões.

Além disso, será explorada a aplicação de algoritmos evolutivos para a seleção de variáveis no processo de predição de alvura. Inspirados na teoria da evolução natural, esses algoritmos utilizam mecanismos como seleção, cruzamento e mutação para explorar o espaço de possíveis conjuntos de variáveis e identificar aqueles que proporcionam o melhor desempenho do modelo. Métodos como Algoritmos Genéticos (GA) serão utilizados para avaliar diferentes combinações de variáveis, permitindo a descoberta de interações complexas que poderiam passar despercebidas em abordagens tradicionais. A aplicação de algoritmos evolutivos pode não apenas otimizar a seleção de variáveis, mas também melhorar a precisão e a generalização do modelo ao explorar um espaço de soluções mais amplo e adaptativo. Isso é particularmente relevante em ambientes industriais dinâmicos, onde as condições operacionais podem variar significativamente, exigindo modelos que sejam capazes de se adaptar de maneira eficaz.

Um possibilidade é a validação do modelo em diferentes unidades industriais, pois a generalização dos resultados obtidos são essenciais para garantir a aplicabilidade e eficácia do modelo em diversos contextos operacionais. Isso pode envolver a adaptação do modelo a diferentes tipos de celulose e processos de branqueamento, garantindo que as técnicas desenvolvidas sejam amplamente aplicáveis na indústria.

Além disso, a aplicação de técnicas de explicabilidade de modelos de aprendizado de máquina (e.g., *SHAP values*, LIME) pode ser explorada para entender melhor como cada variável de entrada impacta as previsões do modelo. Essa compreensão mais profunda pode fornecer *insights* valiosos para a otimização do processo e para a identificação de novas oportunidades de melhoria.

Considerando a análise de componentes principais (PCA) realizada neste trabalho, novos métodos de redução de dimensionalidade ou seleção de características podem ser estudados. Métodos como t-SNE ou UMAP, por exemplo, podem oferecer novas perspectivas sobre os dados e potencialmente melhorar a performance dos modelos preditivos ao lidar com a dimensionalidade de maneira mais eficiente.

Em resumo, os trabalhos futuros se concentrarão na otimização contínua da etapa de pré-processamento e de treinamento do modelo, na integração do preditor com sistemas de controle automatizado, na investigação de novas técnicas de modelagem e na aplicação de métodos avançados de redução de dimensionalidade. Esses esforços visam aprimorar ainda mais a precisão, robustez e aplicabilidade prática dos modelos preditivos no ambiente industrial de branqueamento de celulose.

Referências

CAMPOS, T. K. Instrumento virtual inteligente para previsão de emulsão água/óleo. Instituto Federal de Educação do Espírito SAnto - IFES, Serra, 2022.

CARVALHO, R. D.; LEITE, R. S.; SILVA, H. L. da; SILVA, E. N.; GROSSI, M. A. de A.; COSTA, D. S. da; MOREIRA, R. C.; MORAIS, M. C. G. de; SANTOS, C. M. dos. Redução da sulfidade do licor de cozimento kraft com utilização do sesquisulfato de sódio para regular o ph no estágio d(hot) do branqueamento. São Paulo, Brasil, Outubro 2019.

CAUX, L. S. d.; DALVI, L. C.; AMORIM, S. C. Avaliação de tecnologias para pré e pós-branqueamento visando a produção de polpa branqueada (ecf-light) de eucalyptus urograndis. In: **46th International Pulp and Paper Congress da ABTCP - Associação Brasileira Técnica de Celulose e Papel**. São Paulo, Brasil: [s.n.], 2013.

CHEN, M.; LIU, Q.; CHEN, S.; LIU, Y.; ZHANG, C.-H.; LIU, R. Xgboost-based algorithm interpretation and application on post-fault transient stability status prediction of power system. **IEEE Access**, v. 7, p. 13149–13158, 2019. Disponível em: <https://ieeexplore.ieee.org/document/8620201>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. Association for Computing Machinery, New York, NY, USA, p. 785–794, 2016. Disponível em: <https://doi.org/10.1145/2939672.2939785>.

CHEN, T.; HE, T.; BENESTY, M.; KHOTILOVICH, V.; TANG, Y.; CHO, H. **Xgboost: extreme gradient boosting**. 2015. 1–4 p. R package version 0.4-2.

COLODETTE, J. L.; GOMES, C.; RABELO, M.; EIRAS, K.; OLIVEIRA, K. Branqueamento de polpa kraft de eucalipto e novos desenvolvimentos. **O Papel**, Viçosa, p. 88–111, Setembro 2006.

COLODETTE, J. L.; GOMES, C. M.; RABELO, M.; EIRAS, K. M.; VIÇOSA, M. Progress in eucalyptus kraft pulp bleaching. In: **2nd International colloquium on eucalyptus pulp (2ICEP); Concepcion-Chile**. [S.l.: s.n.], 2005. p. 1–18.

CORREIA, F. M.; D'ANGELO, J. V. H.; JúNIOR, F. G. d. S. Revisitando número kappa: Conceitos e aplicações na indústria de celulose. **O PAPEL**, v. 80, n. 07, p. 77–89, JUL 2019.

FOELKEL, C. Individualização das fibras da madeira do eucalipto para a produção de celulose kraft. eucalyptus online book and newsletter, p. 1–10, 2009.

GOMIDE, J. L. **Curso de Pós-Graduação Lato Senso em Tecnologia de Celulose e Papel**. 2002. 117 p. Apresentação. Características da Polpação, Universidade Federal de Viçosa, Laboratório de Celulose e Papel.

GOMIDE, J. L.; GOMES, F. J. B. Produção e composição das polpas não branqueadas. In: ____. **Branqueamento de polpa celulósica: da produção da polpa marrom ao produto acabado**. Viçosa, MG: Ed. UFV, 2015. cap. Polpas Químicas para Papel.

IBA. [S.l.]: Instituto Brasileiro de Árvores - Dados Estatísticos, 2019. <https://www.iba.org/dados-estatisticos>. Acessado em [21/01/2024].

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. New York: Springer, 2002. Disponível em: <https://link.springer.com/book/10.1007/b98835>.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. 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/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>.

LAROCCA, C. B.; BRITTO, V. R. de; FILHO, E. F. de S.; SANTOS, Y. T.; SILVA, I. C. da; FARIAS, C. T.; ALBUQUERQUE, M. C. Classificação de tensões em chapas de aço if utilizando aprendizado de máquina aplicado a sinais de correntes parasitas pulsadas. XL Simpósio Brasileiro de telecomunicações e processamento de sinais, Santa Rita do Sapucaí, 2022.

LOPES, A. R.; LIRA, J. M. S.; OLIVEIRA, L. A.; GARUZZO, M. d. S. P. B.; BARBALHO, M. V. de S.; ARAúJO, P. O. C. de; SANTOS, G. A. dos; NACIF, J. A. Predição do incremento médio anual volumétrico de eucalyptus com aprendizado de máquina. In: SBC. **Anais do XIV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais**. [S.l.], 2023. p. 81–90.

MACARRINGUE, A. M. J. S. **Fatores de Formação do Preço do Frete Rodoviário de Grãos Agrícolas: Uma Abordagem de um Modelo de Regressão Linear Múltipla por Seleção de Variáveis**. Tese (Doutorado) — Universidade Estadual de Campinas - UNICAMP, 2022.

MATHUR, A.; ANDERSSON, N.; SMITH, D.; ONOFRE, R.; MORGAN, G. Bleach plant optimization utilizing novel measurement technologies complemented with advanced process control. **O Papel**, v. 79, p. 65–72, 02 2018.

NOVEL, B. P. O. U. Measurement technologies complemented with advanced process control. **O PAPEL**, v. 79, n. 2, p. 65–72, 2018.

PAULA, K. G. F. D. **Sensor virtual de alvura em polpa branqueada de celulose baseado em Inteligência Artificial**. 52 p. Dissertação (Trabalho de conclusão de curso) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022. MBA em Inteligência Artificial e Big Data.

PEDRO, V. M. P. **Agentes de Alvura. Avaliar a Influência dos Agentes de Alvura ou Brilho Ótico em Amostras de Substratos para Impressão**. Tese (Doutorado) — Instituto Superior de Educação e Ciências, Julho 2017. Capítulo: Diferentes Procedimentos para Medir o Brilho Ótico e a Brancura - TAPPI Brightness; ISO Brightness; CIE Whiteness.

PROPEQ, P. e Pesquisa em E. Q. **Indústria de papel e celulose: importância e tendências**. 2022. Acessado em: 24 de junho de 2024. Disponível em: <https://propeq.com/industria-de-papel-e-celulose-importancia-e-tendencias/>.

RABELO, M. S. **Tecnologias Avançadas para Pré-Branqueamento de Polpa Kraft de Eucalipto**. 4-16 p. Tese (Doutorado) — Universidade Federal de Viçosa, Viçosa, Minas Gerais, 2006.

RAGNAR, M.; BACKA, S. Hot chlorine dioxide bleaching – a modified approach. **Nordic Pulp & Paper Research Journal**, v. 19, n. 4, p. 417–419, 2004. Disponível em: <https://doi.org/10.3183/npprj-2004-19-04-p417-419>.

REIS, H. M. dos. Processo de extração. In: ____. **Processo de Extração de Celulose Kraft**. 1. ed. [S.l.]: Fontenele Publicações, 2021. p. 40–43.

ROCHA, S. S.; PIANUCCI, M. N.; PITOMBO, C. S.; CUNHA, A. L. Uso de redes neurais para previsão de produção de viagens: Uma análise agregada. In: ASSOCIAçãO NACIONAL DE PESQUISA E ENSINO EM TRANSPORTES (ANPET). **Anais do XXIX Congresso Nacional de Pesquisa e Ensino em Transporte**. [S.l.], 2015. v. 9, p. 1995–2006.

SANDBERG, T. Pdms-comos data transfer development. Metropolia University of Applied Sciences, 2017.

SANTOS, A. C. d. Aprendizado de máquina aplicado na detecção de fraudes em cartão de crédito. 2023.

SANTOS, D. S. Aprendizado de máquina: estatística bayesiana em método de regressão linear simples com aplicação em magnitudes de quasares. 2018.

SANTOS, R.; HART, P. Kraft ecf pulp bleaching: A review of the development and use of techno economic models to optimize cost, performance, and justify capital expenditures. **Tappi Journal**, v. 12, p. 19, 10 2013.

SÖDERHOLM, P.; BERGQUIST, A.-K.; SÖDERHOLM, K. Environmental regulation in the pulp and paper industry: impacts and challenges. **Current Forestry Reports**, Springer, v. 5, p. 185–198, 2019.

VENTORIM, G.; COLODETTE, J. L.; COSTA, M. M. d.; BRITO, A. C. d. Branqueamento ecf e tcf de celulose de fibras secundárias. **Ciência Florestal**, Santa Maria, v. 9, n. 2, p. 41–54, 1999.