UNIVERSIDADE FEDERAL DO MARANHÃO
Coordenação de Engenharia Química/CCET
Trabalho de Conclusão de Curso - TCC

MATEUS GONÇALVES DE MESQUITA

MODELOS PREDITIVOS PARA O CONTROLE DA TURBIDEZ DA ÁGUA EM ETA INDUSTRIAL: UM ESTUDO COMPARATIVO ENTRE REGRESSÃO LINEAR E RANDOM FOREST

São Luís
2025

MATEUS GONÇALVES DE MESQUITA

MODELOS PREDITIVOS PARA O CONTROLE DA TURBIDEZ DA ÁGUA EM ETA INDUSTRIAL: UM ESTUDO COMPARATIVO ENTRE REGRESSÃO LINEAR E RANDOM FOREST

Trabalho de Conclusão de Curso apresentado ao Colegiado de Curso da Engenharia Química do Centro de Ciências Exatas e Tecnologia da Universidade Federal do Maranhão, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia Química.

Orientador: Prof. Dr. Antonio Carlos Daltro de Freitas

São Luís
2025

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Mesquita, Mateus Gonçalves de.
Modelos preditivos para o controle da turbidez da água em ETA industrial: um estudo comparativo entre regressão linear e random forest / Mateus Gonçalves de Mesquita. - 2025.
46 f.

Orientador(a): Antonio Carlos Daltro de Freitas.
Curso de Engenharia Química, Universidade Federal do Maranhão, São Luís, 2025.

1. Turbidez. 2. Pac 18. 3. Modelo Preditivo. 4. Regressão Linear. 5. Floresta Aleatória. I. Freitas, Antonio Carlos Daltro de. II. Título.

BANCA EXAMINADORA:

Prof. Dr. ANTONIO CARLOS DALTRO DE FREITAS
Orientador – DEEQ/UFMA

Profa. Dra. ANNAMARIA DÓRIA SOUZA VIDOTTI
DEEQ/UFMA

Eng. Me. JULLES MITOURA DOS SANTOS JÚNIOR
FEQ/UNICAMP

08 de agosto de 2025

Dedico este trabalho aos meus pais, à minha família e às amizades verdadeiras que me acompanharam nessa longa jornada.

AGRADECIMENTOS

Gostaria de agradecer primeiramente aos meus pais, que me deram suporte emocional e financeiro para que eu pudesse concluir esta jornada. Em todas essas ocasiões, meus pais estiveram comigo e, por isso, sou infinitamente grato: em 2016, quando disse que faria engenharia química; em 2017, quando passei para o curso que queria; em 2019, quando pensei em abandonar a engenharia química para tentar psicologia ou jornalismo (cursos totalmente diferentes da engenharia, por conta da minha frustração); em 2020, quando precisei refazer o ENEM, pois minha situação acadêmica era crítica; em 2021, quando entendi que ser engenheiro químico é, de fato, o que quero para o resto da minha vida; em 2023, quando me mudei para Imperatriz para realizar o meu maior sonho acadêmico: estagiar em uma grande empresa química (Suzano Papel e Celulose); em 2024, quando decidi encerrar meu contrato de estágio para voltar para casa; e em 2025, concluindo este trabalho.

Meus agradecimentos mais sinceros aos meus tios, tias, primos, primas e à minha avó, que sempre confiaram no meu potencial e, ao longo dessa caminhada, sempre se importaram comigo de maneira muito natural e carinhosa.

Sou também infinitamente grato pelo companheirismo daquelas pessoas que rotulei como “amigos de verdade”. Como sou filho único, essas pessoas supriram a figura de irmãos e irmãs. Acertei na escolha de cada um em quem confiei e que andou ao meu lado nesses anos. Sem essas pessoas, eu não teria conseguido, pois me apoiaram em momentos que nem meus pais souberam que existiram, já que eu não queria preocupá-los.

Agradeço muito por ter aparecido, há alguns meses, uma pessoa que vem sendo minha companheira e confidente a cada segundo: minha namorada. Ela tem sido uma enorme fonte de força nos últimos meses, tornando minha rotina conturbada mais suave e fácil de lidar.

Agradeço também aos meus colegas de trabalho que tive em Imperatriz e que acabaram se tornando amigos e família em uma cidade onde eu não conhecia nada nem ninguém, dando-me apoio na vida pessoal e profissional. Sem eles, este estudo, literalmente, não teria sido realizado.

Reservo este espaço também para os meus novos colegas de trabalho da Ambev, por terem me acolhido muito bem e por serem tão compreensivos nos dias em que precisei me ausentar para a conclusão deste trabalho.

Por último, mas não menos importante, gostaria de agradecer a mim mesmo, por nunca ter desistido de mim, ainda que tenha duvidado se conseguiria finalizar este ciclo acadêmico em alguns momentos. Não me dei por vencido.

"O momento é seu, é melhor nunca deixá-lo ir, você só tem uma chance, não perca a chance de estourar, essa oportunidade vem uma vez na vida."

Eminem

MESQUITA, M. G. **Modelos Preditivos para o Controle da Turbidez da Água em ETA Industrial: Um Estudo Comparativo entre Regressão Linear e Random Forest.** 2025. 47 f. Trabalho de Conclusão de Curso de Engenharia Química do Centro de Ciências Exatas e Tecnologia da Universidade Federal do Maranhão, São Luís, 2025.

RESUMO

Este projeto surgiu no contexto de uma estação de tratamento de água (ETA) em uma fábrica de celulose, situada no Maranhão, com o objetivo de substituir a etapa de determinação de dosagem ideal do coagulante, nesse caso o policloreto de alumínio 18% ou PAC 18 (que era feita de forma empírica pelo operador de equipamentos do setor, a partir da sua experiência no trabalho) por um método de determinação da dosagem de coagulante com embasamento estatístico. Para cumprir esse objetivo, decidiu-se criar dois modelos preditivos e compará-los. Um mais tradicional, feito com regressão linear múltipla, e outro mais moderno, utilizando o método da floresta aleatória. Ambos modelos trabalharam com a mesma base de dados contendo 206 observações. O modelo de regressão linear criado utilizou dados de turbidez da água de entrada do sistema (água bruta) e da turbidez da água decantada, atingindo p-valor iguais a 0,000 para ambas as variáveis independentes e R^2 igual a 0,961. Já o modelo de floresta aleatória atingiu um R^2 igual a 0,825, não performando bem durante treinamento e teste nos pontos extremos de dosagem e turbidez. O modelo de regressão linear foi escolhido para ser utilizado no processo e obteve sucesso (turbidez da água decantada abaixo de 1,8 NTU) em 94,92% das ocasiões testadas *in loco*.

Palavras-chave: Turbidez. PAC 18. Modelo Preditivo. Regressão Linear. Floresta Aleatória.

MESQUITA, M. G. **Predictive Models for Water Turbidity Control in Industrial WTP: A Comparative Study between Linear Regression and Random Forest.** 2025. 47 f. Graduate Work (Graduate in Chemical Engineering) – Curso de Engenharia Química do Centro de Ciências Exatas e Tecnologia da Universidade Federal do Maranhão, São Luís, 2025.

ABSTRACT

This research project emerged in the context of a water treatment plant (WTP) in a pulp mill located in Maranhão, with the objective of replacing the stage of determining the ideal dosage of the coagulante, in this case, 18% aluminum polychloride or PAC 18 (which was empirically performed by the equipment operator in the sector, based on their work experience) with a method for determining the coagulant dosage based on statistical foundations. To fulfill this objective, it was decided to create and compare two predictive models. One more traditional, made with multiple linear regression, and another more modern, developed using the random forest method. Both models worked with the same database containing 206 observations. The linear regression model used turbidity data from the system's inlet water (raw water) and the turbidity of the decanted water, achieving p-values equal to 0.000 for both independent variables and an R2 of 0.961. The random forest model, in turn, achieved an R2 of 0.825, not performing well during training and testing at the extreme points of dosage and turbidity. The linear regression model was chosen to be used in the process and was successful (decanted water turbidity below 1.8 NTU) in 94.92% of the tested occasions.

Keywords: Turbidity. PAC 18. Predictive Model. Linear Regression. Random Forest.

SUMÁRIO

1 INTRODUÇÃO

A Estação de Tratamento de Água (ETA) é um conjunto de unidades destinadas à purificação da água bruta captada de uma fonte (como rios ou reservatórios) para torná-la adequada ao consumo humano ou uso industrial. Nas ETAs convencionais são executadas etapas de coagulação, floculação, decantação e filtração, seguidas por desinfecção, com o objetivo de remover impurezas como sólidos suspensos, colóides e microrganismos. Em especial, indústrias de grande porte (como fábricas de celulose) dependem fortemente de suprimento contínuo de água tratada. Nas plantas, a água é usada em processos de produção (lavagem de celulose, geração de vapor, resfriamento etc.) e a qualidade hídrica impacta diretamente a eficiência produtiva. Por isso, o correto funcionamento da ETA é crítico: é comum, por exemplo, que grandes fábricas de celulose invistam em sistemas robustos de tratamento, garantindo água tratada com excelência para suas operações. Essa dependência hídrica torna essencial a automação e o controle preciso das variáveis do processo de tratamento, como o ajuste da dosagem de coagulante para controle da turbidez.

A turbidez da água é uma propriedade física que indica a presença de partículas em suspensão no líquido, sendo uma expressão da opacidade ou “visibilidade reduzida” da água. Em termos técnicos, turbidez é a dispersão e absorção da luz pelas partículas sólidas suspensas na amostra. Essas partículas podem incluir argila, silte, areia, matéria orgânica, algas e diversos microrganismos. Em ambientes naturais, a turbidez apresenta origens distintas: em águas paradas (reservatórios ou lagos), predomina a turbidez devida a partículas coloidais finas, enquanto em águas correntes (rios), eventos de escoamento superficial (especialmente após chuvas intensas) podem remobilizar sedimentos do leito (areia, silte, argila), elevando acentuadamente a turbidez.

A legislação brasileira estabelece que a água tratada para consumo humano tenha turbidez muito baixa (limite de 0,5–1,0 NTU após filtração), seguindo padrões internacionais de segurança (OMS recomenda turvação mediana cerca de 0,1 NTU).

O policloreto de alumínio a 18% (PAC 18) é um coagulante inorgânico amplamente empregado em ETAs industriais devido à sua alta eficiência e estabilidade em faixas largas de pH, além de ser um coagulante com acelerada formação de flocos e maior velocidade de decantação devido à sua estrutura molecular. Comercialmente, o PAC 18 é uma formulação padrão de alta concentração que minimiza a dose necessária e reduz a geração de lodo e custos operacionais. Apesar das vantagens do PAC 18, na fábrica em estudo, a determinação de sua dosagem ideal costumava depender apenas da experiência do operador, sem embasamento

estatístico.

Essa prática subjetiva pode resultar em doses inadequadas: excesso de coagulante gera consumo desnecessário de produto e lodo excessivo, enquanto dose insuficiente deixa a turbidez fora dos padrões desejados. Para eliminar essa incerteza, fez-se necessário um modelo estatístico de controle do processo, que relacionasse propriedades da água bruta (como turbidez, pH, temperatura) com a dose de PAC 18. Modelos estatísticos proporcionam controle do processo ao quantificar como as variáveis de entrada influenciam a saída, permitindo ajustes preditivos e automatizados.

Dentro desse contexto, modelos preditivos de caráter estatístico têm se destacado na indústria para a otimização de variáveis de processo. Em geral, um modelo preditivo utiliza dados históricos do processo para estimar saídas (dose de coagulante) a partir de entradas observadas (turbidez, pH, temperatura). A regressão linear múltipla é uma das abordagens mais tradicionais, em que se ajusta uma equação linear, na qual se estimam coeficientes, que melhor preveem a variável dependente Y em função das variáveis independentes X. O ajuste normalmente ocorre por meio do método dos mínimos quadrados ordinários, que minimiza a soma dos erros quadráticos entre as previsões e os valores observados.

Além da regressão clássica, técnicas de aprendizado de máquina (machine learning) oferecem soluções mais modernas e potentes, especialmente adequadas a situações com muitos dados e relações não lineares. Aprendizado de máquina é um ramo da inteligência artificial que automatiza a construção de modelos analíticos a partir dos dados. De forma simplificada, consiste em algoritmos capazes de aprender padrões e inferir previsões sem serem programados explicitamente para a tarefa específica. Esses métodos iterativos se adaptam continuamente conforme são expostos a novos conjuntos de dados, refinando suas previsões (um aspecto fundamental dado o caráter variável dos insumos industriais). Em termos práticos, o aprendizado de máquina permite modelar relações complexas entre múltiplas variáveis de processo, muitas vezes superando em precisão as técnicas lineares tradicionais quando há interações e efeitos não lineares ocultos.

Na prática industrial, o uso de aprendizado de máquina tem crescido rapidamente no âmbito da Indústria 4.0, devido à disponibilidade crescente de dados e ao poder computacional acessível. Assim, algoritmos preditivos vêm sendo empregados em diversas frentes. Por exemplo, na manutenção preditiva de equipamentos (detectando falhas iminentes antes que ocorram) e no controle de qualidade automatizado (visão computacional identificando produtos defeituosos). Tais aplicações industriais ilustram que soluções de aprendizado de máquina podem aprimorar desde a logística e gestão de estoque até as operações de campo,

contribuindo para reduzir custos e aumentar a eficiência operacional.

Entre os algoritmos de *machine learning* utilizados, a Floresta Aleatória (*Random Forest*) destaca-se pela robustez e versatilidade. Em síntese, florestas aleatórias tendem a entregar alta acurácia preditiva e estabilidade, mesmo quando há ruído nos dados ou relações não-lineares complexas. Dada sua capacidade de manejar múltiplas variáveis explicativas sem requisitos rígidos de distribuição, o algoritmo de *random forest* é considerado uma escolha moderna e eficaz para criar modelos preditivos na indústria.

Poucos estudos brasileiros foram realizados nesse tema específico (modelagem preditiva para tratamento de água e aprendizado de máquina), em que mais esclarecimentos são necessários e ao utilizar de diferentes métodos, um tradicional e um moderno, este estudo contribui para a melhoria do controle de qualidade no tratamento de água industrial, promovendo ganhos de eficiência operacional e impacto econômico positivo, além de contribuir para a comunidade científica brasileira, visando à evolução das tratativas de ocorrências na indústria.

2 OBJETIVOS

O desenvolvimento deste trabalho auxiliará na exclusão da etapa empírica do processo de tratamento da água da fábrica de celulose em estudo, representada pela definição da dosagem ideal de coagulante (PAC 18), criando um modelo preditivo com embasamento estatístico a partir de dados reais coletados durante a rotina de trabalho da fábrica.

Esta pesquisa fornecerá informações para evitar perdas de produto final por turbidez fora de parâmetro de trabalho (PAC 18), bem como evitar uma perda econômica que uma dosagem errada de coagulante pode causar, considerando que tempo, produção e orçamento são pilares essenciais no cotidiano industrial.

2.1 OBJETIVO GERAL

Este estudo tem como objetivo geral desenvolver um modelo preditivo computacional capaz de determinar a dosagem ideal de PAC 18, de acordo com as propriedades da água de entrada da ETA, utilizando regressão linear múltipla e métodos de *machine learning*.

2.1 OBJETIVOS ESPECÍFICOS

- Criar um registro virtual de dados de turbidez;
- Transferir os dados escritos no caderno físico para o registro virtual;
- Criar modelo de regressão linear múltipla;
- Criar modelo de floresta aleatória (*machine learning*);
- Comparar os modelos criados;
- Testar viabilidade dos modelos criados no processo real (testes *in loco*);
- Adaptar o modelo escolhido, caso mais viável, para a rotina operacional;
- Explicar o motivo do procedimento pré-estudo ser inviável, apesar de já utilizado há anos na fábrica;
- Explicar a metodologia e resultados do estudo para os colaboradores do setor em estudo da fábrica;
- Ensinar operadores a utilizarem o modelo escolhido.

3 REVISÃO BIBLIOGRÁFICA

A Estação de Tratamento de Água (ETA) instalada na unidade de produção de celulose, localizada em Imperatriz-MA, tem como principal função fornecer água filtrada com vazão de até 7000 m³/h para uso industrial e 25 m³/h de água potável, a partir da captação no Rio Tocantins (VEOLIA, 2012). Esse processo é fundamental para garantir a qualidade da água utilizada em diferentes setores da planta industrial, sendo a estrutura composta por sistemas como o Multiflo (responsável pela dosagem de químicos, coagulação, floculação e decantação), filtros de areia, tratamento de lodo, entre outros.

O controle da presença de sólidos suspensos e dissolvidos é um dos focos da ETA, que possui uma eficiência de remoção estimada em 96%. O tratamento inclui dosagem de químicos, todos voltados à agregação e sedimentação das partículas (VEOLIA, 2012).

A qualidade da água em mananciais superficiais pode ser profundamente afetada por fatores climáticos e antrópicos. O estudo de Silva et al. (2021) demonstra que o regime de chuvas, aliado ao uso inadequado do solo, exerce influência direta na elevação da turbidez da água, comprometendo seu uso para abastecimento público. Segundo Silva et al. (2021), "a mudança de uso do solo e as consequências para os corpos d'água se tornam ainda mais críticas no contexto de bacias hidrográficas cujos corpos d'água são utilizados para o abastecimento público", reforçando a dificuldade encontrada no controle dos parâmetros de estações de tratamento de água durante períodos de chuvas intensas.

3.1 TURBIDEZ E COAGULANTE

A Resolução CONAMA nº 357/2005 define como "classe 1" as águas doces que, entre outras finalidades, podem ser destinadas ao abastecimento humano após tratamento simplificado e à proteção de comunidades aquáticas. Nessa classe, a turbidez máxima permitida é de 40 NTU, o que reforça a importância do controle rigoroso da matéria particulada em suspensão, especialmente em ambientes lóticos e lênticos (CONAMA, 2005).

Conforme destaca CONAMA (2005), para águas doces de classe 2, que incluem rios utilizados para abastecimento com tratamento convencional, a turbidez pode chegar ao limite de 100 NTU. Esse valor mais elevado, se comparado à classe 1, reflete o reconhecimento da maior carga de partículas e impurezas presentes em cursos d'água com menor exigência de qualidade.

É importante salientar que, segundo a Resolução nº 357/2005, o parâmetro “turbidez” não deve ser analisado isoladamente. Ele compõe um conjunto de condições que, juntas, asseguram que a água não apresente riscos à saúde pública ou comprometa os usos estabelecidos no enquadramento do corpo hídrico (CONAMA, 2005).

Conforme destacam Pinto, Miranda e Pires (2001), a turbidez é uma característica relacionada à presença de partículas em suspensão, como argilas, siltes, matéria orgânica e organismos microscópicos. Essa propriedade interfere diretamente na penetração da luz na coluna d’água e pode afetar tanto processos biológicos quanto o desempenho de sistemas de irrigação.

A água de rios é particularmente suscetível à variação de qualidade ao longo do tempo e do espaço, devido à influência de fatores climáticos, geológicos e antrópicos. Segundo os autores, essa variabilidade torna necessário um monitoramento contínuo para que se possa garantir o uso seguro e eficiente da água nos diferentes setores (PINTO; MIRANDA; PIRES, 2001).

A turbidez elevada pode ser indicativa de processos erosivos na bacia hidrográfica ou da presença de efluentes industriais e domésticos, os quais aumentam significativamente a carga de sólidos em suspensão nos corpos d’água (PINTO; MIRANDA; PIRES, 2001). Tal condição não apenas prejudica o uso da água, como também compromete o equilíbrio dos ecossistemas aquáticos. A avaliação da qualidade da água deve ser feita com base em parâmetros específicos para cada uso pretendido.

A turbidez foi identificada como um indicador ambiental eficiente para detectar alterações na qualidade da água do Rio Tocantins, especialmente durante a estação chuvosa. Isso se deve à mobilização de sedimentos causada pelas chuvas e intensificada pelas atividades agrícolas e mineradoras. A comparação entre os meses seco e chuvoso evidenciou uma diferença estatisticamente significativa nos níveis de turbidez, que saltaram de uma média de $5{,}79 \pm 1{,}09$ NTU no mês seco para $76{,}72 \pm 201{,}24$ NTU no mês chuvoso (SILVA et al., 2021).

Silva et al. (2021) afirmam que os altos índices de turbidez durante o período chuvoso implicam em um aumento dos custos de tratamento da água para consumo humano, uma vez que mais insumos químicos são necessários para torná-la potável.

A turbidez é um parâmetro de grande relevância para a avaliação da qualidade da água, especialmente por estar relacionada à presença de partículas suspensas que interferem na passagem da luz e na estética da água. No entanto, sua medição ainda apresenta obstáculos, principalmente em regiões onde o acesso a equipamentos laboratoriais é limitado (FALCADE; COLOMBO; MANNICH, 2017).

Turbidez elevada tem implicações sérias: além de causar aparência turva indesejável, sólidos em suspensão atuam como vetores de nutrientes (nitrogênio, fósforo), metais pesados (Hg, Pb, Cd, Cu, Zn), toxinas orgânicas (pesticidas, PCBs) e patógenos (FALCADE; COLOMBO; MANNICH, 2017). Essas substâncias podem deteriorar a qualidade do recurso hídrico, causar danos ecológicos (por exemplo, sufocar larvas de peixes no leito) e prejudicar a potabilidade da água.

Rocha (2025) utilizou o ensaio *jar test* como método para avaliar a eficiência de coagulante (quitosana) na remoção da turbidez. Esse ensaio é considerado padrão para testes de coagulação/floculação, permitindo simular, em escala de bancada, as condições de operação em sistemas reais de tratamento. Durante os testes, foi observada uma tendência de maior eficiência da quitosana em pH ácido, com remoções superiores a 95% de turbidez em condições otimizadas. Esses resultados reforçam a influência do pH na performance do coagulante.

Rocha (2025) também empregou técnicas de modelagem estatística para analisar os resultados obtidos, utilizando regressão polinomial de segunda ordem para correlacionar a eficiência de remoção de turbidez com variáveis como pH, dosagem de quitosana e velocidade de agitação. A modelagem estatística foi validada com base no coeficiente de determinação (R^2), que indicou bom ajuste dos dados experimentais ao modelo proposto, além de permitir a identificação de condições ótimas de operação para o processo de coagulação. Esses exemplos reforçam a necessidade de substituir procedimentos empíricos por modelos baseados em dados reais da rotina da ETA, de modo a evitar parâmetros fora de especificação e reduzir perdas econômicas.

De acordo com Cagliari (2018), “o policloreto de alumínio (PAC) é um coagulante amplamente utilizado no Brasil por sua eficiência em diferentes faixas de pH e por apresentar menor geração de lodo”. Essa característica faz do PAC uma escolha recorrente em ETAs que lidam com variações sazonais de qualidade da água bruta. A autora aplicou ensaios de *jar test* para avaliar a eficiência de diferentes dosagens de PAC em comparação com a poliacrilamida. Os resultados obtidos nos experimentos demonstraram que o *jar test* é uma ferramenta indispensável para definir parâmetros operacionais com base em condições reais da ETA.

Cagliari (2018) observou que, durante os testes com PAC, houve redução significativa da turbidez logo após o término da coagulação e floculação, o que reforça o potencial do coagulante para águas com cargas elevadas de sólidos suspensos. Em dosagens controladas, pode otimizar o processo de tratamento, aumentando a eficiência na remoção da turbidez sem comprometer a estabilidade da água tratada

Marques e Campos (2023) observaram que “as melhores remoções de turbidez ocorreram com a aplicação de 20 mg/L de PAC” para água com turbidez até 10 NTU, o que demonstra que a dosagem do coagulante exerce influência direta sobre a eficiência do processo. A variação das doses permitiu identificar um ponto ótimo para o sistema avaliado. A eficiência do PAC foi comprovada experimentalmente com remoções superiores a 90% da turbidez em todas as amostras com dosagens entre 10 e 30 mg/L. Os autores destacam a importância de testes laboratoriais prévios para adequar o processo às características específicas da água bruta.

O estudo de Padilha et al. (2011) comparou a eficiência de três coagulantes – cloreto férrico, sulfato de alumínio e policloreto de alumínio, utilizando o ensaio de *jar test*, sendo este último o que apresentou melhor desempenho. O PAC, mesmo em menor dosagem (10 mg/L), foi capaz de reduzir significativamente a turbidez da água bruta, o que representa vantagem econômica e operacional para estações de tratamento de água. O ensaio *jar test* foi essencial para determinar os parâmetros hidráulicos ideais para os diferentes coagulantes. O método permitiu avaliar gradientes de mistura rápida e lenta, tempos de floculação e velocidades de sedimentação, possibilitando a otimização do processo de coagulação.

Segundo Schmidt (2014), o policloreto de alumínio (PAC) tem se consolidado como uma alternativa eficiente ao sulfato de alumínio no tratamento de água, principalmente por apresentar melhor desempenho em baixas concentrações e menor geração de lodo.

De acordo com Schmidt (2014), “o PAC apresentou melhor desempenho na remoção de turbidez em relação ao sulfato de alumínio em todas as dosagens testadas, sendo mais eficiente principalmente nas faixas de pH mais elevadas”. Essa constatação reforça a flexibilidade operacional do PAC em diferentes condições.

A autora destaca que a dosagem ideal do coagulante depende diretamente da qualidade da água bruta. No caso do PAC, foi observada a melhor eficiência na remoção de turbidez com dosagens entre 25 e 30 mg/L, proporcionando uma turbidez final inferior a 1,0 UNT. O estudo também apontou que a menor necessidade de ajuste de pH para o PAC, comparado ao sulfato de alumínio, contribui para uma operação mais simples e econômica das estações de tratamento (SCHMIDT, 2014).

O estudo de Souza, Souza e Pereira (2015) investigou a aplicação do PAC como coagulante no tratamento de efluente de lavanderia industrial, revelando sua eficácia na remoção de cor, turbidez e demanda química de oxigênio (DQO). Os ensaios foram realizados com *jar test*, utilizando dosagens entre 0,1 e 0,6 mL/L, e indicaram remoções superiores a 94% da turbidez nas concentrações a partir de 0,2 mL/L.

A análise estatística dos dados foi realizada por meio do teste de Tukey, que indicou não haver diferença significativa entre as dosagens de 0,20 a 0,60 mL/L para os parâmetros avaliados. Isso demonstra uma estabilidade na performance do PAC, mesmo com pequenas variações na dosagem (SOUZA; SOUZA; PEREIRA, 2015).

3.2 ESTATÍSTICA E MODELOS PREDITIVOS

A análise estatística de dados tem como objetivo extrair padrões relevantes e evidências quantitativas que sustentem conclusões científicas. Gelman, Hill e Vehtari (2021) afirmam que o foco principal da estatística aplicada deve estar em entender as variáveis envolvidas e a estrutura dos dados antes da aplicação de qualquer modelo matemático. A regressão linear é apresentada pelos autores como uma ferramenta essencial para modelar relações entre variáveis quantitativas, facilitando a interpretação de fenômenos e a previsão de resultados futuros.

De acordo com Gelman, Hill e Vehtari (2021), o coeficiente de determinação R^2 pode não ser um bom resumo do ajuste do modelo, pois pode ser alto mesmo em modelos mal especificados. Os autores sugerem que, apesar de ser amplamente utilizado, o R^2 deve ser interpretado com cautela e em conjunto com outras métricas. Os autores também fazem críticas ao uso indiscriminado de p-valores como critério de decisão em análises estatísticas. Segundo eles, os p-valores são úteis para resumir informações, mas não fornecem uma resposta absoluta sobre se um efeito é real ou não.

Essa visão reforça a necessidade de uma abordagem mais contextual e menos mecânica na avaliação de significância de uma variável em um modelo de regressão, que não deve ser avaliada apenas com base em valores-p, mas sim considerando a plausibilidade dos efeitos estimados, os intervalos de confiança e o conhecimento prévio do fenômeno estudado (GELMAN; HILL; VEHTARI, 2021).

A análise estatística de dados requer uma abordagem criteriosa que combine conhecimento do problema com ferramentas matemáticas adequadas. Harrell Jr. (2015) afirma que o valor de um modelo estatístico está menos na sua complexidade e mais na sua capacidade de fornecer inferências estáveis e interpretáveis. No contexto da regressão linear, o autor defende a modelagem como um processo contínuo e iterativo, que deve considerar suposições, transformação de variáveis e validação interna.

Harrell Jr. (2015) é crítico quanto ao uso do coeficiente de determinação R^2 como medida única de desempenho: “R^2 tem utilidade limitada, especialmente ao comparar modelos,

pois não leva em conta o sobreajuste ou a complexidade do modelo" Ele recomenda a adoção de métricas de validação cruzada e penalização para evitar interpretações equivocadas.

A significância estatística de uma variável deve ser acompanhada de uma análise de relevância prática. Harrell Jr. (2015) defende que o foco deve ser deslocado do simples "significante ou não significante" para a avaliação do tamanho do efeito, variabilidade dos dados e plausibilidade científica.

A interpretação de p-valores também é problematizada pelo autor. Segundo ele, "pequenos p-valores são frequentemente usados como evidência contra a hipótese nula, mas seu uso indevido leva a uma falsa sensação de certeza" (HARRELL JR., 2015). Ele sugere que intervalos de confiança e análises de sensibilidade oferecem informações mais robustas.

A análise estatística é apresentada por Lu (2021) como uma disciplina que vai além da aplicação de fórmulas, exigindo compreensão teórica das estruturas matemáticas subjacentes. O autor defende que um modelo estatístico rigoroso deve ser construído com base na teoria de probabilidade, garantindo validade inferencial.

A regressão linear é formalizada como um modelo de expectativa condicional da variável resposta, dado um conjunto de covariáveis. Segundo Lu (2021), esse modelo fornece uma estimativa ótima no sentido dos mínimos quadrados ordinários (OLS), quando as suposições do modelo são atendidas.

Lu (2021) afirma que "R^2 mede a proporção da variância na variável de resposta que é explicada pelo preditor linear". Apesar de sua popularidade, o autor alerta que essa métrica pode ser enganosa em modelos com muitas variáveis irrelevantes, pois tende a aumentar artificialmente com a inclusão de termos no modelo.

Sobre o p-valor, Lu (2021) explica que ele representa a probabilidade de se observar um valor estatístico tão extremo quanto o observado, sob a suposição de que a hipótese nula é verdadeira. Ele ressalta, no entanto, que a interpretação desse valor deve ser contextualizada e acompanhada de outras métricas, como os intervalos de confiança.

A significância de uma variável, no contexto da regressão, deve ser entendida como uma evidência probabilística de que seu coeficiente não é nulo. Contudo, Lu (2021) adverte que "significância estatística não implica relevância prática", sugerindo que o tamanho do efeito e a variabilidade são igualmente importantes na análise.

A análise estatística de dados é uma ferramenta essencial para transformar dados brutos em informações úteis à tomada de decisão. Fávero e Belfiore (2017) destacam que a estatística aplicada tem papel central em estudos empíricos nas ciências sociais, econômicas e administrativas, permitindo o teste de hipóteses e a construção de modelos explicativos.

A regressão linear é apresentada como um dos métodos mais importantes da estatística inferencial, pois permite identificar e quantificar relações entre variáveis (FÁVERO; BELFIORE, 2017).

De acordo com os autores, “o coeficiente de determinação R^2 mede o grau de explicação da variabilidade da variável dependente proporcionado pelo modelo” (FÁVERO; BELFIORE, 2017). No entanto, eles ressaltam que R^2 não deve ser interpretado isoladamente, sendo necessário avaliar a significância dos coeficientes e a validade das suposições do modelo.

Fávero e Belfiore (2017) explicam que o p-valor “indica o grau de significância estatística de uma variável explicativa, sendo utilizado para testar a hipótese de que seu coeficiente seja estatisticamente diferente de zero”. Assim, valores de p inferiores a 0,05 indicam evidência contra a hipótese nula, ao nível de 5% de significância.

Os autores alertam que a significância estatística de uma variável não deve ser confundida com relevância prática. É possível encontrar coeficientes estatisticamente significantes, mas com impacto irrelevante na variável dependente. Por isso, Fávero e Belfiore (2017) recomendam avaliar o tamanho dos coeficientes conjuntamente com os testes de significância.

A modelagem estatística para previsão da dosagem de coagulante com base em parâmetros da água bruta pode otimizar o processo de coagulação em ETAs, reduzindo desperdícios e melhorando a qualidade do tratamento. Leite et al. (2023) propõem um modelo de regressão linear múltipla com variáveis como turbidez, cor aparente e pH. A metodologia empregada no estudo seguiu etapas clássicas da análise de regressão: verificação de multicolinearidade, análise de resíduos, identificação de *outliers* e validação cruzada do modelo. Os autores destacam que o modelo final foi validado com base no coeficiente de determinação ajustado ($R^2 = 0,931$) e no teste F global, que indicaram sua robustez estatística. Também foi observado que os resíduos do modelo apresentaram distribuição aproximadamente normal e ausência de padrão nos gráficos de resíduos *versus* valores ajustados, o que indica que os pressupostos do modelo linear foram atendidos de forma satisfatória.

Segundo os autores, “o modelo de regressão linear ajustado indicou que a turbidez foi a variável com maior influência sobre a dosagem de coagulante” (LEITE et al., 2023). Essa evidência reforça a importância de monitorar continuamente esse parâmetro na água bruta como critério de controle operacional. Os autores recomendam que o modelo seja utilizado como ferramenta de apoio à tomada de decisão nas estações de tratamento de água, podendo ser atualizado periodicamente com novos dados operacionais para manter sua acurácia preditiva.

Conforme apontado por Sávio (2023) da KPMG, a combinação de *big data* com técnicas de *machine learning* é uma das aplicações mais importantes da automação industrial contemporânea. Esse arranjo "permite que as máquinas aprendam por si mesmas a partir de uma enorme quantidade de dados, aperfeiçoando sua capacidade de tomada de decisões".

A inteligência artificial (IA) é definida como "o estudo de agentes que recebem percepções do ambiente e executam ações" (RUSSELL; NORVIG, 2020). A obra destaca que o objetivo central da IA é projetar agentes racionais, capazes de agir de forma eficaz mesmo diante da incerteza, tornando-se especialmente relevantes em contextos complexos e dinâmicos.

No campo do aprendizado de máquina, os autores explicam que o foco é "construir programas de computador que melhorem automaticamente com a experiência" (RUSSELL; NORVIG, 2020). Tal definição conecta diretamente a IA com a capacidade de generalização a partir de dados, o que é fundamental para aplicações em larga escala.

Russell e Norvig (2020) afirmam que "as árvores de decisão são classificadores altamente interpretáveis que dividem os dados com base nos atributos mais informativos em cada nó". Essas estruturas são amplamente utilizadas pela sua simplicidade e facilidade de visualização, o que favorece sua aplicação em contextos onde a explicabilidade é essencial.

Quanto às florestas aleatórias, os autores descrevem que "*Random forests* são conjuntos de árvores de decisão geradas a partir de subconjuntos aleatórios dos dados e dos atributos, cuja decisão final é obtida por votação" (RUSSELL; NORVIG, 2020). Essa abordagem melhora a precisão da classificação ao reduzir o risco de *overfitting* presente em árvores individuais.

As florestas aleatórias (*random forests*) são descritas como "um dos métodos de aprendizado supervisionado mais poderosos e amplamente utilizados na prática moderna de mineração de dados" (BIAU; SCORNET, 2015). Essa popularidade se deve tanto ao seu desempenho preditivo quanto à sua robustez contra *overfitting*.

Segundo os autores, uma floresta aleatória "é uma combinação de árvores de decisão construídas usando subconjuntos aleatórios do conjunto de dados e subconjuntos aleatórios das variáveis" (BIAU; SCORNET, 2015). O modelo final é obtido pela agregação dos resultados individuais das árvores, normalmente por votação majoritária no caso de classificação.

Biau e Scornet (2015) explicam que as árvores de decisão, que formam a base das florestas aleatórias, "são algoritmos de particionamento recursivo que dividem o espaço de entrada em regiões homogêneas". Essas divisões são feitas com base nos atributos que melhor separam os dados em cada ponto de decisão.

No que se refere à quantidade mínima de dados necessária para o uso eficaz de florestas aleatórias, os autores observam que “a consistência estatística dos modelos de floresta aleatória depende da taxa de crescimento do número de árvores e do tamanho da amostra” (BIAU; SCORNET, 2015). Isso implica que amostras pequenas podem levar a modelos instáveis, a menos que se controle o número de divisões por árvore e o nível de aleatoriedade.

Os autores também destacam que “embora as florestas aleatórias não sejam algoritmos fáceis de interpretar, sua performance na prática é notável, mesmo com pouca afinação dos hiperparâmetros” (BIAU; SCORNET, 2015), o que contribui para sua ampla adoção em aplicações de aprendizado de máquina.

Sobre o tamanho da amostra, James et al. (2013) observam que “florestas aleatórias são particularmente eficazes quando há um grande número de preditores, mesmo que o número de observações não seja muito alto”. Ainda assim, eles destacam que conjuntos de dados muito pequenos podem comprometer a estabilidade do modelo, especialmente se houver muitas variáveis irrelevantes.

Em termos comparativos, os autores afirmam que “a regressão linear tem um desempenho inferior em problemas com fortes interações e não linearidades, onde métodos como *random forest* geralmente obtêm resultados superiores” (JAMES et al., 2013). Assim, a escolha do modelo depende do tipo de relação entre os preditores e a variável resposta.

Segundo Kuhn e Johnson (2013), “os modelos de floresta aleatória podem ser extremamente eficazes em muitas aplicações práticas e requerem pouca parametrização por parte do usuário”. Os autores destacam que, mesmo com conjuntos de dados de tamanho moderado, “as florestas aleatórias tendem a produzir bons resultados, desde que o número de variáveis preditoras seja razoável e a quantidade de ruído não seja excessiva”. Isso sugere que não há um número mínimo fixo de observações, mas sim uma relação entre a complexidade dos dados e a estabilidade do modelo.

Em comparação com a regressão linear, Kuhn e Johnson (2013) apontam que “a regressão linear é limitada quando as relações entre as variáveis não são lineares ou quando há muitas interações importantes entre os preditores”. Nesses casos, métodos baseados em árvores, como *random forest*, apresentam desempenho superior.

Segundo Probst, Wright e Boulesteix (2018), “a floresta aleatória pode produzir bons resultados mesmo em conjuntos de dados com alto número de variáveis e relativamente poucas observações, desde que os hiperparâmetros sejam escolhidos de maneira apropriada”. Isso sugere que a qualidade do ajuste depende tanto da estrutura dos dados quanto do processo de *tuning*.

Ao comparar florestas aleatórias com modelos lineares, os autores ressaltam que “modelos lineares podem falhar em capturar interações complexas entre variáveis, enquanto as florestas aleatórias lidam bem com esses aspectos, muitas vezes sem necessidade de pré-processamento” (PROBST; WRIGHT; BOULESTEIX, 2018).

Em relação à parametrização, os autores explicam que “os principais hiperparâmetros que influenciam o desempenho de uma floresta aleatória são o número de árvores, o número de variáveis consideradas em cada divisão, e a profundidade máxima das árvores” (PROBST; WRIGHT; BOULESTEIX, 2018). A escolha inadequada desses parâmetros pode comprometer tanto a acurácia quanto a generalização do modelo.

Com isso, tem-se a bibliografia para o estudo realizado, abordando diversos autores, em seus diferentes tópicos. Entretanto, percebe-se a estabilidade de opiniões entre os autores, tanto para a turbidez e coagulantes, que afirma a melhor eficiência do PAC-18 perante outros coagulantes durante o tratamento de água e efluentes, quanto para os modelos preditivos, que explicitam as limitações da regressão linear e a capacidade de desenvolvimento do *random forest*.

4 MATERIAL E MÉTODOS

4.1 MATERIAL

Como material base tem a água captada a partir do Rio Tocantins, chamada de água bruta. Ela possui gases dissolvidos (gás carbônico, oxigênio, amônia, nitrogênio etc.), sólidos dissolvidos (sais de cálcio e magnésio, cloretos, alcalinidade, sulfatos etc.), sólidos suspensos (lama, sujeiras, material orgânico e microbiológico, areia etc.). Essas impurezas presentes influenciam o pH (ficando entre 6 e 8), turbidez, cor aparente, alcalinidade e condutividade da água.

O presente estudo utilizou-se como coagulante policloreto de alumínio 18%, comumente chamado de PAC 18, possuindo como fórmula geral $Al_n(OH)_mCl_{(3n-m)}$ e "18%" sendo referência a concentração de óxido de alumínio (Al_2O_3) na sua composição, definindo assim a sua basicidade. Ele é um líquido viscoso amarelado, no qual suas cadeias poliméricas pré-hidroxiladas geram espécies insolúveis que neutralizam cargas elétricas de partículas suspensas (orgânicas, inorgânicas, metais). Produz flocos mais densos, tornando sua decantação mais rápida que coagulantes tradicionais (como sulfato de alumínio), acelerando a sedimentação e facilitando a remoção de impurezas. Hidrolisa-se lentamente, mantendo eficácia em amplas faixas de pH (4 a 9) sem necessidade de ajuste frequente de alcalinidade, porém sua faixa ótima de trabalho está entre 6,5 e 6,8 pH, removendo assim turbidez, cor, matéria orgânica e microrganismos.

Outros dois materiais utilizados foram a soda cáustica (NaOH), para o aumento do pH e ácido clorídrico (HCl) para a redução do pH da água bruta. O NaOH foi utilizado por possuir alta solubilidade em água e o HCl foi utilizado por ser um ácido mineral e barato.

4.2 LOCAL DA PESQUISA

O estudo foi desenvolvido em uma Estação de Tratamento de Água (ETA) de uma fábrica de celulose situada na cidade de Imperatriz (MA), às margens do Rio Tocantins. Nesse local, durante o período de coleta de dados, o clima variou entre sol intenso e chuvas fortes, com a temperatura ambiente ficando entre 20 °C e 35 °C. O local da captação da água bruta é uma região movimentada por barcos menores pesqueiros e barcos maiores turísticos e de fiscalização.

4.3 INSTRUMENTOS/EQUIPAMENTOS/SOFTWARES UTILIZADOS

4.3.1 Fase Laboratorial

Foi utilizado um turbidímetro portátil da marca Fast Tracker, com faixa de medição de 0,00 a 9,99; 10,0 a 99,9 e de 100 a 1000 NTU e resolução de 0,01 NTU de 0,00 a 9,99 NTU; 0,1 NTU de 10,0 a 99,9 NTU; 1 NTU de 100 a 1000 NTU.

Usou-se um pHmetro de bancada da marca Bel Engeneering, faixa de trabalho entre 0 e 14 pH e resolução de 0,01 pH por toda a faixa. Ele possui calibração de temperatura e de solução tampão automática.

Como equipamento de teste de jarro (*jar test)*, foi utilizado um sistema automático da marca Ethik Technology de 6 jarros simultâneos. O sistema é equipado com um dosador que permite a aplicação simultânea de coagulantes e polímeros entre os jarros, além de contar com um mecanismo que viabiliza a retirada simultânea de amostras da água clarificada, contemplando diferentes taxas de sedimentação. Possui ainda um dispositivo destinado à coleta simultânea das amostras da fase superior (sobrenadante), garantindo a precisão das informações obtidas e a consistência na comparação dos resultados.

4.3.2 Fase Computacional

Na fase computacional do estudo foi utilizado o Google Colaboratory, comumente conhecido como Google Colab, por ser uma plataforma de computação em nuvem, baseada em navegadores web para programação em python, o qual já vem pré-configurado com uma ampla gama de bibliotecas já instaladas, incluindo *NumPy*, *Pandas*, *Matplotlib* e *Scikit-learn*. Além da capacidade de executar código em servidores remotos do Google, o que significa que os usuários podem realizar computações intensivas mesmo em dispositivos com recursos computacionais limitados, como tablets ou computadores antigos, desde que tenham acesso à internet.

4.3.2.1 Bibliotecas do Python

Para o desenvolvimento das análises estatísticas e dos modelos preditivos apresentados neste trabalho, foram utilizadas diversas bibliotecas da linguagem de programação python. Tais bibliotecas são amplamente reconhecidas na comunidade científica

e são fundamentais para o tratamento de dados, construção de modelos e avaliação de desempenho. A seguir, são descritas as principais bibliotecas utilizadas, conforme sua aplicação.

A biblioteca Pandas foi utilizada para a manipulação e estruturação dos dados tabulares. Por meio de sua principal estrutura, o *DataFrame*, foi possível realizar operações como leitura de arquivos, filtragem de registros, agrupamentos, transformações e junções de dados. Sua aplicação foi essencial para o preparo adequado das bases de dados utilizadas nas etapas analítica e preditiva.

Para as análises estatísticas e modelagem linear, utilizou-se a biblioteca *Statsmodels*, voltada para métodos estatísticos clássicos. Essa biblioteca foi fundamental na construção de modelos de regressão linear e na obtenção de métricas como p-valores, intervalos de confiança e testes de hipóteses, permitindo uma avaliação detalhada do comportamento das variáveis e da significância dos coeficientes estimados.

A representação gráfica dos dados e resultados foi realizada por meio da biblioteca *Matplotlib*, que possibilitou a criação de gráficos como histogramas, diagramas de dispersão e curvas de tendência. A visualização dos dados foi essencial tanto na etapa exploratória quanto na apresentação dos resultados, facilitando a compreensão dos padrões e relações entre as variáveis analisadas.

Para avaliação quantitativa do desempenho dos modelos, foram utilizadas métricas fornecidas pela biblioteca *Scikit-learn* por meio de seu módulo de métricas. A métrica aplicada (coeficiente de determinação) foi fundamental para mensurar a precisão das previsões geradas pelos modelos, além de auxiliar na comparação entre abordagens distintas.

Ainda utilizando a biblioteca *Scikit-learn*, foi empregada uma função para dividir o conjunto de dados em subconjuntos de treinamento e teste. Essa separação é uma prática essencial em projetos de modelagem preditiva, pois permite validar o desempenho do modelo com dados independentes, contribuindo para uma avaliação mais realista da sua capacidade de generalização.

Por fim, foi utilizado o algoritmo *RandomForestRegressor*, um método de aprendizado de máquina do tipo ensemble, baseado na construção de múltiplas árvores de decisão. Esse algoritmo realiza a agregação dos resultados das árvores para gerar previsões mais robustas, reduzindo a variância do modelo e mitigando o risco de sobreajuste. Sua aplicação foi importante para modelar relações complexas entre as variáveis e alcançar maior acurácia nos resultados.

4.4 ANÁLISES, MÉTODOS E PROCEDIMENTOS EXPERIMENTAIS

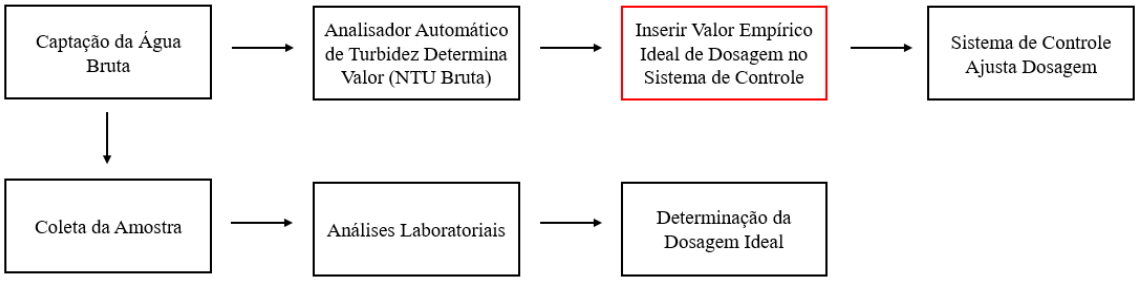
4.4.1 Visão Geral do Procedimento de Dosagem de PAC 18

A água do Rio Tocantins era captada e na entrada da ETA era feita a coleta da amostra para análise dos parâmetros de controle em laboratório, porém ela também passava pelos medidores automáticos de vazão, turbidez, pH e temperatura.

A leitura da turbidez era feita pelo turbidímetro automático inserido no processo determinando o valor em NTU da turbidez da água vinda do rio. Com esse valor, o operador responsável por inserir no sistema de controle a dosagem de PAC 18 (a partir do seu conhecimento adquirido durante a sua rotina de trabalho, de forma empírica) determinava o valor ideal de dosagem de PAC 18 e inseria esse valor no sistema de controle. Simultaneamente, eram realizadas as análises laboratoriais, que demoravam uma média de 50 minutos para serem finalizadas.

Caso o valor de dosagem de coagulante que o operador julgou como ideal tivesse fornecido uma turbidez da água decantada dentro do limite de trabalho (1,8 NTU), ou seja, caso o operador acertasse o valor de dosagem, o processo seguia como estava. A Figura 1 explica esse processo de tomada de decisão, com acerto do operador na dosagem observacional.

Figura 1- Fluxograma caso o operador acertasse a dosagem de coagulante.



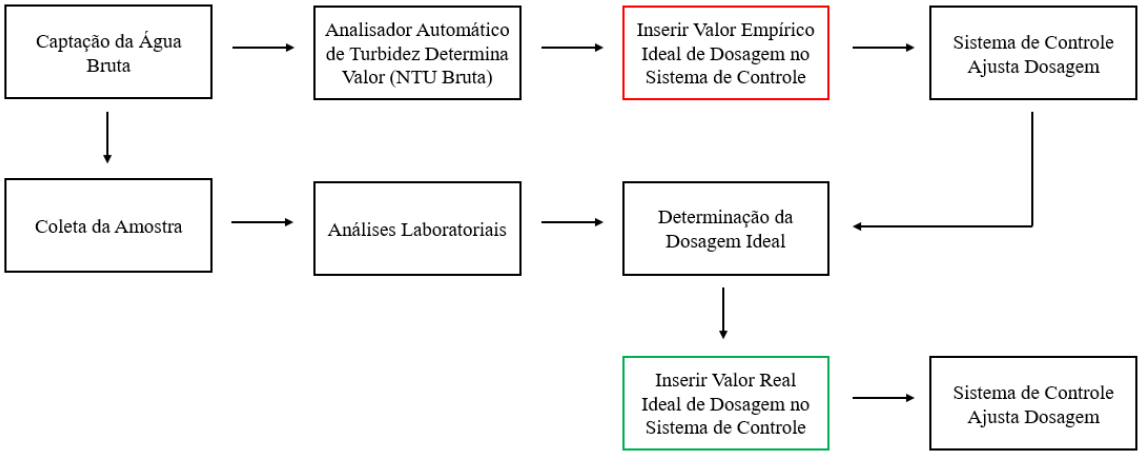
Fonte: Próprio Autor (2025).

O valor da turbidez da água decantada era determinado por um turbidímetro automático também, após a água passar pelo sistema de decantação. Da leitura da turbidez da água bruta pelo turbidímetro automático até a leitura da turbidez da água decantada passam-se, aproximadamente, 40 minutos.

A análise de laboratório serve como um “plano B”, um apoio caso o operador não acerte a dosagem ideal de coagulante. Ao sair a leitura de turbidez da água decantada e ser

confirmado que ela estava fora do limite de trabalho, era necessário esperar a análise de laboratório determinar a dosagem ideal e inserir esse valor no sistema de controle. Por fim, espera-se mais 40 minutos para o turbídimetro automático na saída do sistema determinar a nova turbidez da água decantada. Se a turbidez ainda ficar acima de 1,8 NTU, repetem-se as análises laboratoriais. O processo em caso de erro na dosagem pelo mostrador é apresentado de forma descritiva na Figura 2.

Figura 2- Fluxograma caso o operador errasse a dosagem de coagulante.



Fonte: Próprio Autor (2025).

4.4.1.1 Registro dos Dados

Os dados de turbidez da água bruta, pH, temperatura, dosagem ideal de PAC 18 e a turbidez da água decantada fornecida pela dosagem de PAC 18 inserida no sistema de controle eram registradas em um caderno físico, que tinha que ser preenchido manualmente a cada turno.

4.4.2 Procedimento de *Jar Test*

O ensaio de teste de jarro é realizado em bancada para simular os processos de coagulação e floculação, estimando a dose mínima de coagulante necessária para atingir os objetivos de tratamento da água. Neste caso, emprega-se o coagulante PAC 18 e como amostra a água bruta captada do Rio Tocantins. A água é coletada na entrada da ETA e levada para o laboratório de análises que fica logo ao lado, iniciando-se imediatamente as medições.

Inicialmente, mede-se a temperatura, o pH e a turbidez da água, utilizando termômetro, pHmetro e turbidímetro, respectivamente. O pH deve ser corrigido para a faixa

entre 6,5 e 6,8 pH se estiver fora, utilizando NaOH ou HCl, dependendo do caso. Em seguida, preparam-se os seis jarros de vidro (preenchidos igualmente com 1 L de amostra cada e posicionados nos agitadores do teste de jarro).

O jarro 1 serve como amostra de referência, não sendo adicionado coagulante. Procede-se com a adição do PAC 18 nos jarros 2 a 6. As doses de PAC são definidas em função da turbidez inicial: quanto maior a turbidez maior a dose requerida. Assim, testa-se uma faixa crescente de dosagens, aumentando a concentração em 1 mg/L (ppm) de jarro a jarro. Espera-se que doses em torno de 18 a 22 mg/L de PAC 18 abranja, na grande maioria dos casos, uma faixa de turbidez entre 6 e 15 NTU na água bruta. Logo, partindo de uma dose inicial estimada de cerca de 18 mg/L no segundo jarro, adicionam-se 19, 20, 21 e 22 mg/L nos jarros subsequentes.

Aciona-se o agitador em velocidade rápida (120 RPM) por 60 segundos, assegurando mistura eficaz do coagulante na amostra. Em seguida, reduz-se a rotação para 30 RPM e mantém-se a agitação lenta por 15 minutos, o que previne a quebra dos flocos formados, favorecendo o crescimento e a sedimentação. Após esse período, as pás agitadoras são levantadas e permite-se que o conteúdo dos jarros decante em repouso por 20 minutos.

Concluído o tempo de sedimentação, coleta-se cuidadosamente o sobrenadante de cada jarro e mede-se a turbidez residual com o turbidímetro. A dose ótima de PAC 18 será escolhida entre os jarros em que a turbidez seja menor ou igual a 1,8 NTU, levando em consideração a estratégia da companhia no período (tópico abordado na seção “4.4.1.1 Dosagem ótima de PAC 18”).

Caso nenhuma das dosagens testadas (jarros 2 a 6) atinja o critério de turbidez menor ou igual a 1,8 NTU, repete-se o procedimento alterando a faixa de dosagem de PAC 18 entre os jarros 2 e 6.

4.4.2.1 Dosagem ótima de PAC 18

Caso apenas um jarro apresentasse turbidez menor ou igual a 1,8 NTU, utilizava-se no processo a dosagem de coagulante do jarro em questão. Porém, na maioria das análises ocorridas durante o período do estudo, ocorreu mais de um jarro com turbidez menor ou igual a 1,8 NTU. Nesses casos, levou-se em consideração a estratégia da fábrica no período da análise. Há dois cenários estratégicos, o primeiro no qual a fábrica pretende gastar menos coagulante, por questões orçamentárias. O segundo cenário, em que se pretende deixar o fator econômico em segunda prioridade, para assim obter-se uma menor turbidez na água decantada.

No primeiro cenário, escolheu-se entre os jarros com turbidez menor ou igual a 1,8 NTU, o jarro que recebeu a menor dosagem de PAC 18, já no segundo cenário, escolheu-se o jarro que recebeu a maior dosagem de coagulante.

4.4.3 Regressão Linear

A regressão linear tem como cerne oferecer uma relação entre uma ou mais variáveis lineares. No presente estudo trabalhou-se com uma regressão linear múltipla (mais de uma variável independente), tendo a equação (1) como equação geral.

$$Y_i = b_0 + b_1 X_{1i} + \cdots b_k X_{ki} + u_i \quad (1)$$

Y é a variável dependente, b_n são os coeficientes, X sendo as variáveis independentes, n representando o número de variáveis independentes, enquanto i representa cada observação feita (ou seja, neste estudo, cada análise feita) e u representa o erro (ou resíduo) entre a variável dependente real e a gerada pelo modelo.

Como variável depende este estudo teve a dosagem de PAC. Suas variáveis independentes disponíveis para análise foram a turbidez da água bruta, pH de trabalho (pH dentro do Multiflo, após adição de ácido ou base), temperatura de trabalho (temperatura dentro do Multiflo), por último, a turbidez da água decantada (turbidez após a água ter passado pelo sistema). Por praticidade, a variável dependente e as variáveis independentes foram nomeadas como PAC, NTU Bruta, pH, Temp e NTU Decantada, respectivamente, na base de dados gerada e nos códigos escritos.

4.4.3.1 Mínimos Quadrados Ordinários

A equação (1), por conter o termo de erro, indica que estamos tratando da equação para obtenção do Y real. Portanto, o Y estimado será a equação (2).

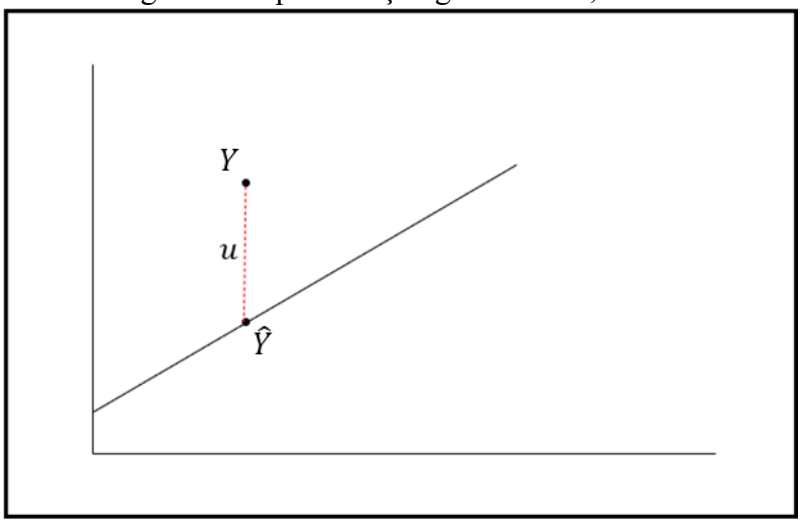
$$\hat{Y}_i = b_0 + b_1 X_{1i} + \cdots + b_k X_{ki} \quad (2)$$

Logo, o termo u, referente ao erro ou resíduo, assume o caráter:

$$u_i = Y_i - \hat{Y}_i \tag{3}$$

Com isso estabelecido, tem-se a base para o método de estimação do modelo de regressão linear por mínimos quadrados ordinários (MQO, que em inglês seria *ordinary least squares* ou OLS). A representação gráfica é apresentada na Figura 3.

Figura 3- Representação gráfica de u, $\hat{Y}$ e Y.



Fonte: Próprio Autor (2025).

Uma das principais intenções com a obtenção de um modelo é minimizar os erros entre a variável dependente real e a estimada. Porém, se apenas for igualada a soma dos erros encontrados a 0, encontram-se alguns problemas, como, por exemplo, a possível existência de mais de uma reta que satisfaça essa condição, além de também não fazer distinção entre a magnitude dos erros menores e dos erros maiores. Sendo assim, surge uma solução: deve-se encontrar a menor somatória possível entre os quadrados dos erros.

$$\min \sum_{i=1}^{n} (u_i)^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{4}$$

A solução da equação (4) é feita derivando-a parcialmente para cada coeficiente (a, b_1, ..., b_k) e igualando a 0. A derivação parcial em b_k, para visualização, é encontrada na equação (5).

$$\frac{\partial \sum_{i=1}^{n} (u_i)^2}{\partial b_k} = \sum_{i=1}^{n} -2 \cdot X_{ki} \cdot (Y_i - \widehat{b_0} - \widehat{b_1} X_{1i} - \cdots - \widehat{b_k} X_{ki}) = 0 \tag{5}$$

Após essa etapa, aplicando a derivação parcial e igualando a 0 para cada coeficiente, teríamos os valores de $\hat{a}$, $\widehat{b_1}$, ..., $\widehat{b_k}$, que são os estimadores de cada coeficiente ou os estimadores de mínimos quadrados ordinários, ficando com a equação (6) sendo a equação do modelo de regressão linear múltipla para o nosso problema.

$$\hat{Y} = \widehat{b_0} + \widehat{b_1}X_1 + \cdots + \widehat{b_k}X_k \quad (6)$$

Para o modelo de regressão linear múltipla criado precisou-se determinar quais variáveis independentes fariam parte do modelo. Por exemplo, o modelo podia ter todas as variáveis independentes, apenas três das quatro, apenas duas ou apenas uma variável independente. Por existirem 4 variáveis independentes, existiam 15 combinações possíveis e diferentes para o modelo. Porém, antes de ser feito o teste de p-valor e R^2, estabeleceu-se que o modelo necessitava conter a variável de turbidez da água bruta, pois era nela que a dosagem de PAC 18 iria agir, além de conter a variável de turbidez da água decantada, pois o limite de 1,8 NTU era algo alterável, dependendo das diretrizes da fábrica. Além disso, foi estabelecido que um modelo que contém apenas uma variável independente era arbitrário demais. Com esses postulados estabelecidos, restaram apenas 4 possíveis combinações, essas combinações são apresentadas na Tabela 1.

Tabela 1: Combinações possíveis e diferentes testadas de variáveis independentes.

Ensaios	Combinações de Variáveis
1	NTU Bruta, NTU Decantada, pH, Temp
2	NTU Bruta, NTU Decantada, pH
3	NTU Bruta, NTU Decantada, Temp
4	NTU Bruta, NTU Decantada

Fonte: Próprio Autor (2025).

4.4.4 Aprendizado de Máquina

O aprendizado de máquina (*machine learning* em inglês) é um ramo de aplicação da inteligência artificial, que tem como objetivo o desenvolvimento de algoritmos e modelos estatísticos que permitem aos sistemas computacionais realizar tarefas específicas sem o uso de instruções explícitas, baseando-se em padrões e inferências a partir de uma base de dados fornecida sem a necessidade de codificar regras fixas.

Um exemplo atual são as recomendações de produtos pelas lojas online, onde o algoritmo da loja aprende com as buscas e compras dos clientes, fazendo novas recomendações

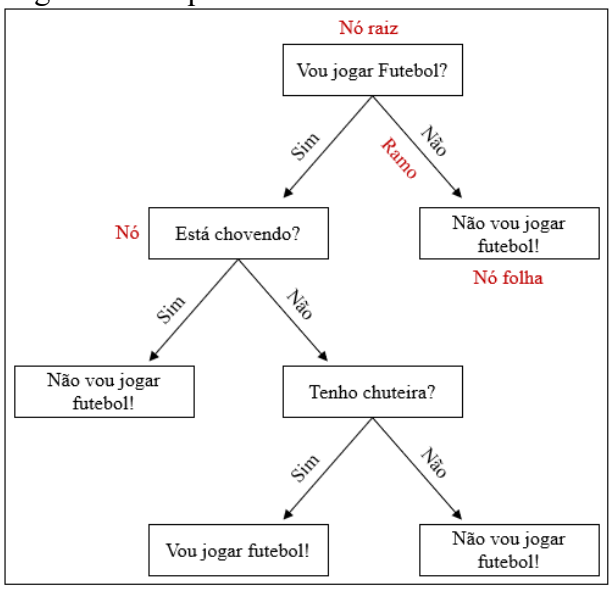
de forma autônoma, sem precisar um ser humano dar algum comando. Quanto mais compras e buscas o cliente faz, mas o algoritmo fica preciso nas recomendações.

Neste estudo foi utilizado uma categoria específica de aprendizado de máquina, chamada de aprendizado supervisionado, no qual o algoritmo aprende a partir de exemplos rotulados, em que tanto as entradas quanto as saídas corretas são fornecidas. O objetivo é encontrar uma função que generalize bem para dados não vistos. Isso inclui problemas de classificação (saídas são categorias discretas, como, por exemplo, a resposta ser "azul" ou "não azul") e regressão (saídas são valores contínuos como, por exemplo, a resposta ser o valor da dosagem de PAC 18 necessária para atingir a turbidez desejada).

4.4.4.1 Árvore de Decisão

Uma árvore de decisão é um fluxograma que, a partir de uma pergunta inicial, a qual se deseja responder, faz uma série de novas perguntas, a partir das possíveis respostas para a pergunta inicial, até chegar à decisão final. A Figura 4 apresenta o esquema da árvore de decisão.

Figura 4 – Esquema de uma árvore de decisão.



Fonte: Próprio Autor (2025).

Por exemplo, uma árvore de decisão para decidir ir jogar ou não futebol, em que a pergunta inicial seria "Vou jogar futebol?", tendo as possíveis respostas "sim" ou "não". Caso "não", gera a decisão "não irei jogar futebol" para a pergunta inicial, o que não gera novas ramificações. Porém, caso "sim", gera uma nova pergunta, por exemplo, "está chovendo?", ramificando para novas respostas possíveis e diferentes, que podem gerar uma decisão final ou uma nova pergunta.

A pergunta inicial pode ser chamada de “nó raiz”, cada nova pergunta gerada pode ser chamada de “nó”, cada resposta para as perguntas pode ser chamada de “ramo” e cada decisão pode ser chamada de “nó folha”, por isso o nome “árvore de decisão”.

4.4.4.1.1 Métrica de Impureza

Em árvores de decisão, uma métrica de impureza serve para quantificar o quão heterogêneo está o conjunto de valores da variável dependente em um determinado nó. Quanto maior a impureza, mais variáveis ou valores diferentes existem naquele nó e quanto menor, mais os alvos são semelhantes entre si. O objetivo do algoritmo é, a cada divisão, escolher o atributo (e o ponto de corte) que produza a maior redução de impureza, ou seja, que torne os nós seguintes os mais homogêneos possíveis em relação ao valor que se quer prever.

Em árvores de decisão para problemas de regressão, o erro quadrático médio ou MSE (*Mean Squared Error*) é a métrica de impureza mais utilizada para avaliar a qualidade de uma divisão em um nó. No modelo criado utilizando aprendizado de máquina neste estudo, MSE foi a métrica de impureza utilizada, pois é a métrica que o módulo da biblioteca importada no código em python utiliza de forma implícita.

4.4.4.2 Floresta Aleatória

Uma floresta aleatória é um método de aprendizagem de conjunto (*ensemble learning*) que opera construindo uma grande quantidade de árvores de decisão (chamado de hiperparâmetro, que é ajustável na codificação em python e neste trabalho foram utilizados 1000 estimadores) que trabalham simultaneamente e sendo independentes uma das outras. Após as árvores fazerem suas previsões, para regressão, o modelo calcula a média dos valores previstos de todas as árvores.

Um conceito importante quando tratando-se de floresta aleatória é o conceito de conjunto de treinamento. Quando criado o modelo, precisa-se dividir a base de dados que se tem entre o conjunto de treinamento e o conjunto de teste. O conjunto de treinamento é usado para ensinar o modelo a reconhecer padrões e fazer previsões, enquanto o conjunto de teste é usado para avaliar a precisão e a generalização do modelo em dados novos e não vistos. A divisão mais comum é utilizar 70% dos dados para treinamento e 30% para teste, que foi a divisão utilizada neste estudo.

Uma floresta aleatória combina duas ideias principais para garantir que as árvores

sejam descorrelacionadas e, assim, o *ensemble* seja mais eficaz. O primeiro é a amostragem com reposição (*bagging* ou *bootstrap aggregating*), que é o processo central do método. Para um conjunto de dados de treinamento com N amostras, o *bagging* cria k novos conjuntos de treinamento (um para cada uma das k árvores da floresta). Cada um desses novos conjuntos também tem tamanho N, mas é criado por meio de amostragem com reposição do conjunto original, que é um método de amostragem onde cada elemento selecionado da população é devolvido ao conjunto original antes da próxima seleção. Isso significa que cada novo conjunto de dados pode conter duplicatas de algumas amostras e omitir outras. Os dados do novo conjunto são selecionados, repostos ao conjunto original e o ciclo se repete.

A segunda etapa da floresta aleatória é a seleção aleatória de atributos para cada nó de cada árvore, com isso, se você possui um número p de atributos totais, apenas um número m (em que m << p, geralmente para regressão, m é um terço de p) de atributos será utilizado em cada nó, com isso, atributos irrelevantes têm menor chance de afetar todas as árvores, assim como uma única variável dominante distorça todas as previsões.

4.5 ANÁLISE ESTATÍSTICA

4.5.1 Coeficiente de Determinação

Para comparação entre os dois modelos, foi escolhida a métrica do coeficiente de determinação ou R^2. O coeficiente de determinação mede a qualidade do ajuste do modelo. R^2 indica o quanto das flutuações em Y o modelo consegue capturar. Por exemplo, R^2 igual a 0,65 significa que 65% da variação de Y (variável dependente) é explicada pelos preditores no modelo, e o restante 35% é ruído ou não explicado. Em outras palavras, reflete o poder explicativo do modelo. R^2 tem como equação:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(u_i)^2} \quad (7)$$

Em que $\bar{Y}$ é a média dos valores de Y real e as outras variáveis permanecem com os significados já apresentados neste estudo.

O R^2 é calculado automaticamente pelos modelos em python com a utilização das bibliotecas já existentes.

4.5.1 Valor-p ou P-valor

No caso da regressão linear, é necessário determinar quais variáveis independentes possuem significância estatística a ponto de fazerem parte do modelo ou serem desconsideradas. Para isto, utilizou-se o valor-p de cada variável independente.

Para cada variável preditora, testou-se se seu coeficiente é significativamente diferente de zero. A hipótese nula (H_0) assume que o coeficiente é zero (nenhum efeito), enquanto a hipótese alternativa (H_1) assume que o coeficiente não é zero. Esse teste é feito a partir da estatística t, que tem a seguinte equação característica:

$$t_{\widehat{b_k}} = \frac{\widehat{b_k}}{erro\ padrão\ (\widehat{b_k})} \tag{8}$$

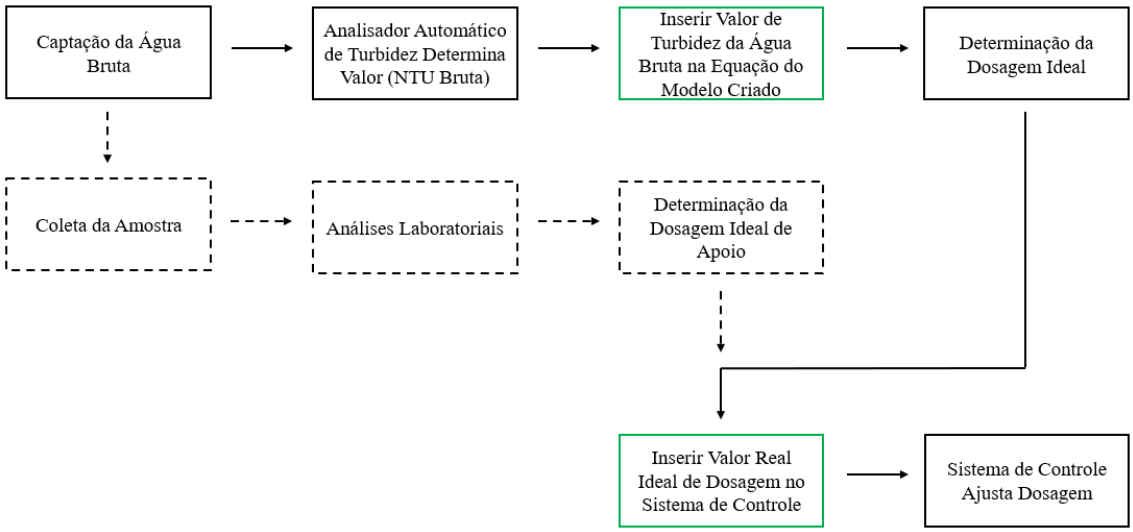
Quanto maior o módulo de t mais improvável que o coeficiente $\widehat{b_k}$ seja igual a 0. O valor-p de $\widehat{b_k}$ é calculado pelo software utilizado como a probabilidade de observar um valor do módulo de $t_{\widehat{b_k}}$ maior ou igual ao calculado. Caso o valor-p seja menor ou igual a 0,05, a variável é significativa para o modelo. Porém, muito importante ressaltar que o valor-p indica a significância estatística de uma variável em um modelo de regressão linear, mas não serve como indicador do impacto dessa variável independente na variável dependente, devendo-se levar em consideração a aplicação prática do modelo criado e a real significância de cada coeficiente.

5 RESULTADOS E DISCUSSÃO

5.1 FLUXOGRAMA DE DOSAGEM DE PAC 18

Houve uma alteração no fluxograma do procedimento de dosagem do PAC 18, esse é apresentado na Figura 5.

Figura 5 – Novo fluxograma do procedimento de dosagem PAC 18.



Fonte: Próprio Autor (2025).

A partir da criação do modelo, o turbidímetro automático em linha faz sua leitura da turbidez da água bruta, o operador utiliza esse o valor na equação do modelo criado, gerando a dosagem ideal de PAC 18. As análises de laboratório seguem ocorrendo simultaneamente e servindo como apoio para o método principal. Caso o modelo criado forneça uma turbidez da água decantada acima do limite de 1,8 NTU, utiliza-se a dosagem encontrada em laboratório.

5.2 REGISTRO DE DADOS

Ao todo, foram feitas 493 observações de dados, durante 164 dias, 3 vezes por dia (três turnos). Entretanto, foram descartadas 287 observações por causa de sinistros, por exemplo, falhas operacionais humanas (manobra operacional feita erroneamente) e mecânicas (quebra de equipamentos), além de períodos de mudanças climáticas muito abruptas (chuva forte, seguida de sol intenso), pois a turbidez da água bruta ficava muito irregular (atingia valores muito altos, seguida de valores baixos), causando leituras não confiáveis do

turbidímetro automático em linha, viabilizando apenas a análise feita em laboratório. Portanto, restaram 206 observações.

Após o início do estudo, o registro de dados passou a ser feito virtualmente, em planilha criada em Excel e armazenada na pasta pública de rede, facilitando o acesso e prevenindo sinistros como avarias ao caderno físico ou, simplesmente, perdê-lo, algo que já aconteceu na fábrica. O registro permaneceu sendo 3 vezes ao dia, uma vez por turno, promovendo um constante aumento na base de dados para análises futuras.

5.3 MODELO CRIADO EM REGRESSÃO LINEAR

As quatro combinações de variáveis foram testadas e comparadas a partir do p-valor das variáveis em cada combinação e do R^2 gerado pelo modelo de cada combinação (Figura 6 a Figura 9). O cálculo de R^2 e p-valor de cada combinação foram feitos implicitamente pelo código em python.

A partir das iterações feitas, foram testadas a combinação 3 (turbidez da água bruta, turbidez da água decantada e temperatura) e combinação 4 (turbidez da água bruta e turbidez da água decantada) para implementação do modelo. Ambos modelos apresentaram variáveis com p-valor abaixo de 0,05 e R^2 acima de 0,95.

A equação (9) representa a combinação 3 de variáveis independentes.

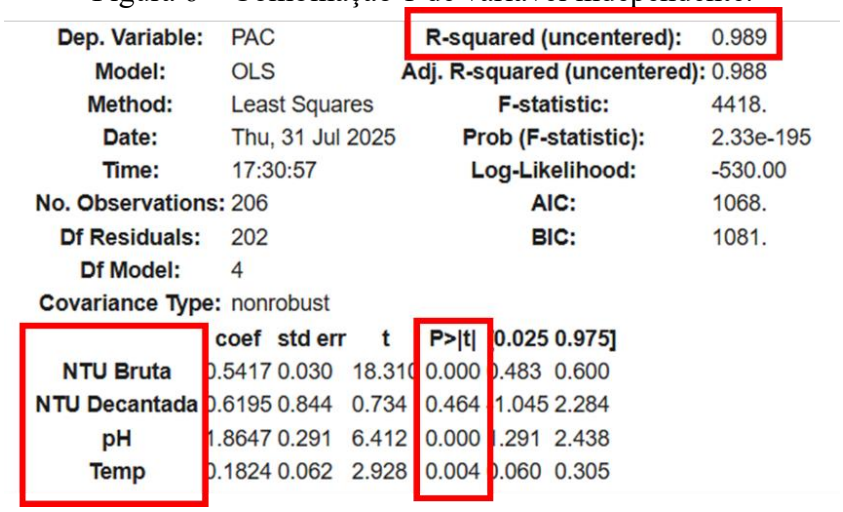
$$PAC = 0{,}5109 \cdot \ NTU\ Bruta + 1{,}9605 \cdot NTU\ Decantada + 0{,}5462 \cdot Temp \qquad (9)$$

Enquanto a equação (10) representa a combinação 4 de variáveis independentes.

$$PAC = 0{,}3275 \cdot \ NTU\ Bruta + 16{,}1459 \cdot NTU\ Decantada \qquad (10)$$

Primeiramente, foi feita uma comparação com as observações adquiridas durante a coleta de dados. Foram inseridos os valores das variáveis independentes observadas nas equações (9) e (10), com exceção da turbidez da água decantada que foi travada em 1,8 NTU, pois é o limite de trabalho. Após isso, foram comparados os valores da dosagem de PAC real que foi feita e a gerada pelo modelo levando em consideração a turbidez da água decantada gerada pela dosagem real.

Figura 6 – Combinação 1 de variável independente.

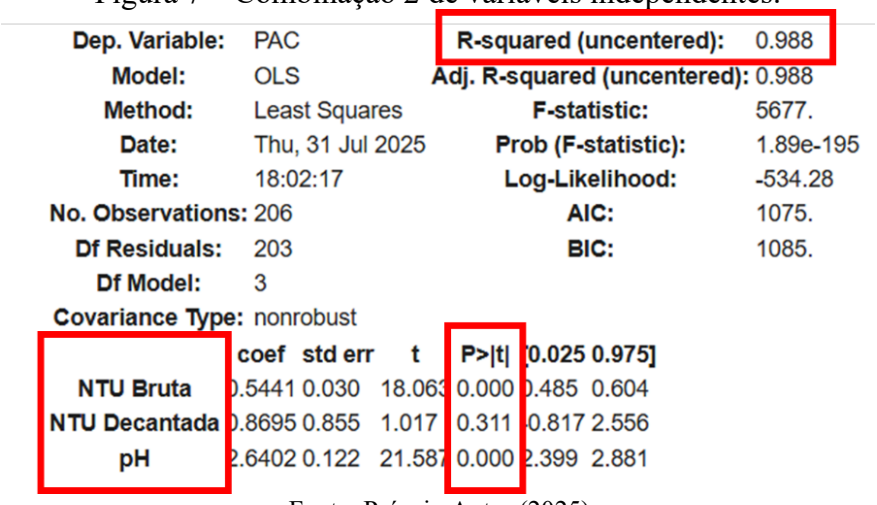


Dep. Variable:	PAC	R-squared (uncentered):	0.989
Model:	OLS	Adj. R-squared (uncentered):	0.988
Method:	Least Squares	F-statistic:	4418.
Date:	Thu, 31 Jul 2025	Prob (F-statistic):	2.33e-195
Time:	17:30:57	Log-Likelihood:	-530.00
No. Observations:	206	AIC:	1068.
Df Residuals:	202	BIC:	1081.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
NTU Bruta	0.5417	0.030	18.310	0.000	0.483	0.600
NTU Decantada	0.6195	0.844	0.734	0.464	-1.045	2.284
pH	1.8647	0.291	6.412	0.000	1.291	2.438
Temp	0.1824	0.062	2.928	0.004	0.060	0.305

Fonte: Próprio Autor (2025).

Figura 7 – Combinação 2 de variáveis independentes.

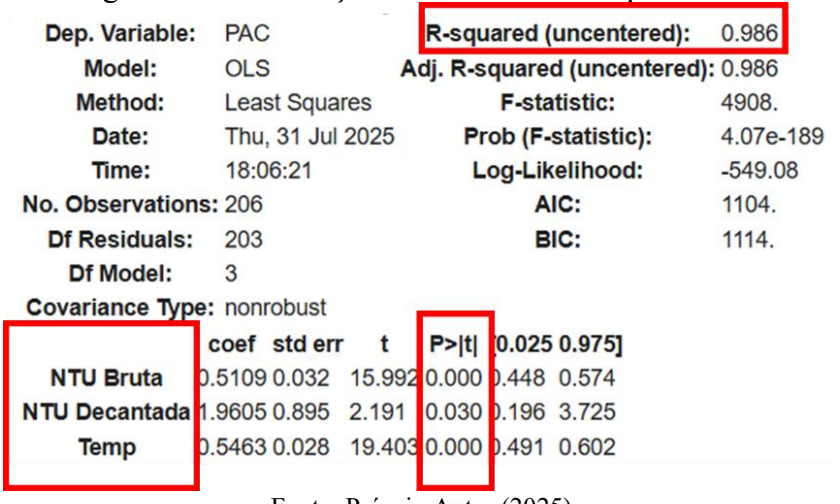


Dep. Variable:	PAC	R-squared (uncentered):	0.988
Model:	OLS	Adj. R-squared (uncentered):	0.988
Method:	Least Squares	F-statistic:	5677.
Date:	Thu, 31 Jul 2025	Prob (F-statistic):	1.89e-195
Time:	18:02:17	Log-Likelihood:	-534.28
No. Observations:	206	AIC:	1075.
Df Residuals:	203	BIC:	1085.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
NTU Bruta	0.5441	0.030	18.063	0.000	0.485	0.604
NTU Decantada	0.8695	0.855	1.017	0.311	-0.817	2.556
pH	2.6402	0.122	21.587	0.000	2.399	2.881

Fonte: Próprio Autor (2025).

Figura 8 – Combinação 3 de variáveis independentes.

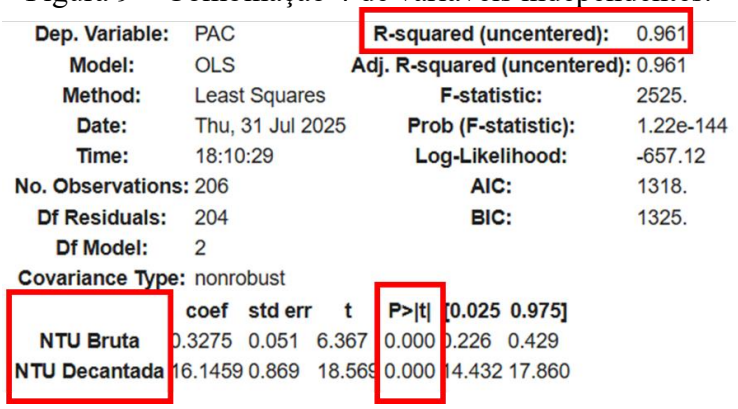


Dep. Variable:	PAC	R-squared (uncentered):	0.986
Model:	OLS	Adj. R-squared (uncentered):	0.986
Method:	Least Squares	F-statistic:	4908.
Date:	Thu, 31 Jul 2025	Prob (F-statistic):	4.07e-189
Time:	18:06:21	Log-Likelihood:	-549.08
No. Observations:	206	AIC:	1104.
Df Residuals:	203	BIC:	1114.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
NTU Bruta	0.5109	0.032	15.992	0.000	0.448	0.574
NTU Decantada	1.9605	0.895	2.191	0.030	0.196	3.725
Temp	0.5463	0.028	19.403	0.000	0.491	0.602

Fonte: Próprio Autor (2025).

Figura 9 – Combinação 4 de variáveis independentes.



Dep. Variable:	PAC	R-squared (uncentered):	0.961
Model:	OLS	Adj. R-squared (uncentered):	0.961
Method:	Least Squares	F-statistic:	2525.
Date:	Thu, 31 Jul 2025	Prob (F-statistic):	1.22e-144
Time:	18:10:29	Log-Likelihood:	-657.12
No. Observations:	206	AIC:	1318.
Df Residuals:	204	BIC:	1325.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
NTU Bruta	0.3275	0.051	6.367	0.000	0.226	0.429
NTU Decantada	16.1459	0.869	18.569	0.000	14.432	17.860

Fonte: Próprio Autor (2025).

Foram estabelecidas 4 situações para determinar se o modelo se comportou melhor, igual ou pior que a observação real:

- Se a turbidez real da água decantada tiver sido maior que 1,8 NTU e a dosagem de PAC do modelo for maior que a real, o modelo performou melhor;
- Se a turbidez real da água decantada tiver sido menor ou igual a 1,8 NTU e a dosagem de PAC do modelo for menor que a real, o modelo performou pior;
- Se a turbidez real da água decantada tiver sido maior que 1,8 NTU e a dosagem de PAC do modelo for menor que a real, o modelo performou pior;
- Qualquer outra situação o modelo se comportou igual ao real.

A partir disso, se o modelo performou melhor ou igual a dosagem de PAC 18 real, foi considerado como um impacto positivo do modelo gerado. Esses dados são compilados e apresentados na Tabela 2 para o modelo gerado pela combinação 3 e na Tabela 3 para o modelo gerado pela combinação 4.

Tabela 2: Comportamento do modelo gerado pela combinação 3.

Comportamento	Contagem de Comportamento
Melhor	16
Igual	111
Pior	79

Fonte: Próprio Autor (2025).

Tabela 3: Comportamento do modelo gerado pela combinação 4.

Comportamento	Contagem de Comportamento
Melhor	20
Igual	172
Pior	14

Fonte: Próprio Autor (2025).

A combinação 3 obteve impacto positivo (soma dos comportamentos iguais e melhores, divididos pelo total de observações) em 61,65% das 206 observações. A combinação 4 obteve 93,2% de impacto positivo. O modelo da equação (10), teoricamente, comportou-se melhor que o modelo da equação (9), porém, ambos foram testados na prática. Durante o primeiro semestre de 2024 foram feitos 177 testes com cada modelo no processo real em que o modelo da equação (9) gerou 70,06% dos valores de turbidez da água decantada abaixo do limite de trabalho, enquanto o modelo da equação (10) gerou 94,92% dos valores de turbidez abaixo de 1,8 NTU, conforme apresentado na Tabela 4.

Tabela 4: Observações abaixo de 1,8 NTU na água decantada durante aplicação real.

Modelo	**Contagem de Observações**	**%**
Equação (9)	124	**70,06%**
Equação (10)	168	**94,92%**

Fonte: Próprio Autor (2025).

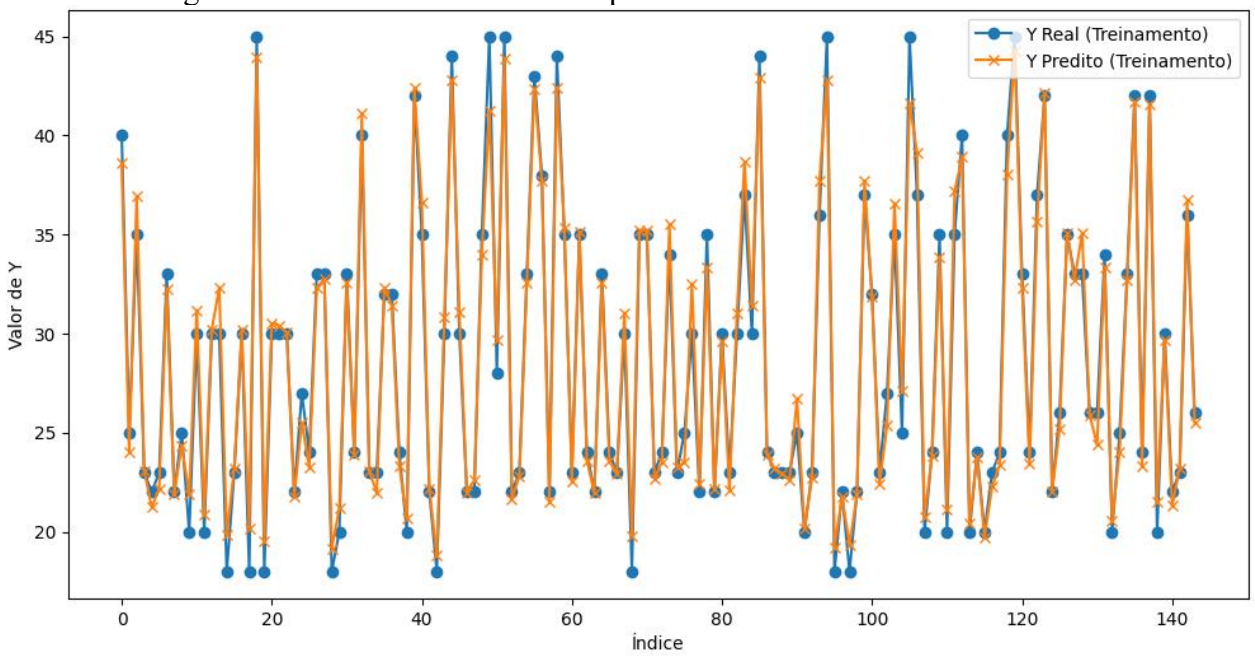
Dessa forma, pode-se concluir que a equação (10) performou melhor tanto na teoria quanto na prática, sendo a mais robusta dentre as testadas.

5.4 MODELO CRIADO EM FLORESTA ALEATÓRIA

Ao contrário da regressão linear, foram utilizadas todas as variáveis possíveis para a elaboração do modelo de floresta aleatória por conta do seu método de análise já explicados neste trabalho. O cálculo de R^2 foi realizado implicitamente pelo modelo e os gráficos foram gerados pelo código escrito em python. Este modelo obteve um R^2 igual a 0,825 (dados apresentados na Figura 11). Notou-se que, nos pontos extremos de dosagem de coagulantes (tanto altos quanto baixos), o modelo não performou bem. Entretanto, nos pontos de dosagem média, o modelo performou bem. É possível verificar a discrepância observada entre os dados calculados e as medidas reais na Figura 12.

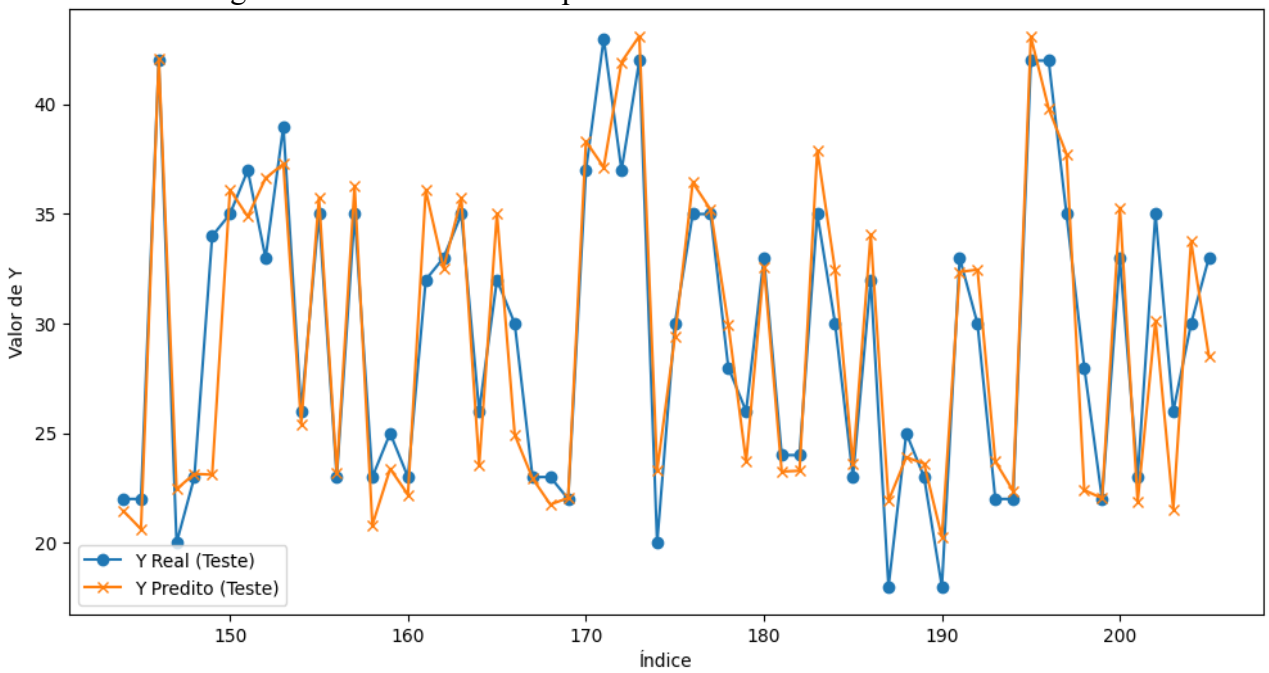
Durante o teste o modelo manteve o comportamento do treinamento, com correção considerável nos pontos extremos. A dispersão do teste obteve um resultado satisfatório. Entretanto, por a base de dados ser relativamente pequena para este modelo, a floresta aleatória teve performance limitada.

Figura 10 – Treinamento realizado pelo modelo de floresta aleatória.



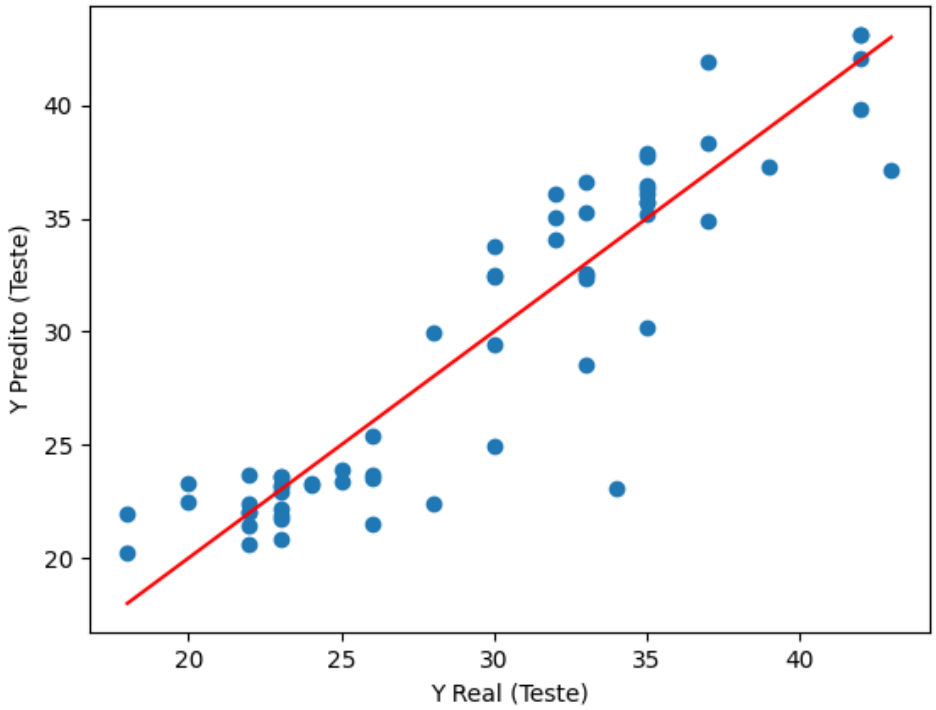
Fonte: Próprio Autor (2025).

Figura 11 – Teste realizado pelo modelo de floresta aleatória.



Fonte: Próprio Autor (2025).

Figura 12 – Diagrama de dispersão do teste realizado pelo modelo de floresta aleatória.



Fonte: Próprio Autor (2025).

Pelo baixo R^2, baixa performance nos pontos de dosagem extrema e pelos modelos de regressão linear múltipla terem apresentado desempenhos satisfatórios, o modelo criado com aprendizado de máquina não chegou a ser testado em processo. Entretanto, o modelo de floresta aleatório tende a ter performance diretamente proporcional com o aumento constante do banco de dados que foi estabelecido pelo estudo. A população utilizada no desenvolvimento desse modelo, aqui apresentado, é considerado ainda um número pequeno, de forma que, para maiores populações de dados, é esperado que o modelo de *Machine Learning* venha a superar a capacidade de predição de dados apresentada pelo modelo de regressão anteriormente apresentado. Essa condição estabelece que no futuro, provavelmente, seja mais adequado utilizar a floresta aleatória, por conta da sua maior robustez ao tratar a base de dados.

Embora modelos baseados em *Random Forest* sejam geralmente reconhecidos por sua capacidade de capturar relações complexas e não lineares entre variáveis, seu desempenho pode ser comprometido em situações com conjuntos de dados pequenos e/ou com baixa variabilidade estrutural. No presente caso, a base de dados utilizada para modelagem do sistema de medida de turbidez da água conta com apenas 260 amostras, o que representa uma limitação substancial para modelos de aprendizado mais complexos.

O modelo de *Random Forest* tende a particionar os dados em subconjuntos (árvores de decisão) e, ao fazer isso com uma quantidade limitada de amostras, pode sofrer com sobreajuste (*overfitting*), capturando ruídos e flutuações locais da base de treinamento, em vez de identificar tendências generalizáveis. Esse problema agrava-se caso o sistema de medição de turbidez possua uma relação predominantemente linear entre as variáveis de entrada (como intensidade de luz, comprimento de onda, concentração de sólidos) e a variável de saída (nível de turbidez).

Por outro lado, a regressão linear, sendo um modelo mais simples e com baixa variância, tende a se beneficiar de conjuntos de dados pequenos e de relações lineares ou quase lineares. Sua estrutura permite uma maior capacidade de generalização nesse tipo de cenário, o que pode resultar em melhores métricas de desempenho, especialmente quando se avaliam indicadores como erro médio quadrático (RMSE) ou coeficiente de determinação (R^2) em validação cruzada.

Além disso, a alta interpretabilidade do modelo linear facilita a identificação e o controle de variáveis influentes, o que pode ser vantajoso no desenvolvimento e calibração de sistemas de medição físicos, como sensores ópticos aplicados à análise de turbidez *in loco*, facilitando e o processo e melhorando a operação industrial.

6 CONCLUSÃO

Ao fim do estudo foi estabelecido o modelo de regressão linear múltipla da equação (10), como o modelo definitivo e a ser utilizado para determinar a dosagem ideal de PAC 18 durante o processo da estação de tratamento de água da fábrica de celulose em trabalho. Com isso, atingiu-se o cerne deste projeto de pesquisa que era a exclusão da etapa empírica do processo.

Além disso, atingiu-se os objetivos específicos de criação de modelos diferentes e comparação de metodologias estatísticas mais tradicionais e metodologias que introduzissem a maior capacidade descritiva e com maior rigor computacional ao processo, utilizando as técnicas de *Machine Learning* usando *Random Florest*.

Os resultados obtidos demonstraram que, apesar da capacidade do modelo *Random Forest* em capturar relações complexas, a regressão linear apresentou melhor desempenho para o conjunto de dados analisados. Isso se deve principalmente ao número reduzido de amostras (260) e à possível relação linear entre as variáveis envolvidas na medição da turbidez da água. Modelos mais simples, como a regressão linear, tendem a generalizar melhor em cenários com poucos dados e menor complexidade, evitando o sobreajuste observado em modelos mais robustos.

A aplicação *in loco* da equação desenvolvida demonstrou a viabilidade de uso operacional, otimizando o processo e minimizando custos com insumos e garantindo a qualidade operacional.

REFERÊNCIAS

BIAU, G.; SCORNET, E. ***A Random Forest Guided Tour.*** arXiv preprint, 2015. Disponível em: arXiv:1511.05741.

CAGLIARI, Larissa. **Padronização do uso de policloreto de alumínio e poliacrilamida em uma ETA de Porto Alegre.** Trabalho de Conclusão de Curso (Graduação) – Escola de Engenharia, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2018.

CONAMA. **Resolução nº 357/2005.** Rio de Janeiro: CONAMA, 2005.

FALCADE, Dóris; COLOMBO, Geovana; MANNICH, Michael. **Tubo de turbidez para determinação de baixo custo da turbidez em corpos d'água superficiais.** Revista de Gestão de Água da América Latina, v. 14, n. 0, 2017.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®.** Rio de Janeiro: Elsevier, 2017.

GELMAN, Andrew; HILL, Jennifer; VEHTARI, Aki. ***Regression and Other Stories.*** Cambridge; New York: Cambridge University Press, 2021.

HARRELL JR., Frank E. ***Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.*** 2. ed. New York: Springer-Verlag, 2015.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. ***An Introduction to Statistical Learning: with Applications in R.*** New York: Springer, 2013.

KUHN, Max; JOHNSON, Kjell. ***Applied Predictive Modeling.*** New York: Springer, 2013.

LEITE, B. J.; ZIMMERMANN, N. E. K.; FERREIRA, V. L. D.; DIAS, F. L. **Modelagem da dependência do consumo de coagulante dos parâmetros brutos da água: ajuste de um modelo de regressão linear.** Scientia Plena, Aracaju, v. 19, n. 11, 2023.

LU, Jun. ***A rigorous introduction to linear models.*** arXiv preprint, 2021. Disponível em: arXiv:2105.04240.

MARQUES, Diego Gouveia; CAMPOS, Valquíria de. **Avaliação de policloretos de alumínio com relação à remoção de turbidez de água bruta da represa Cachoeira do França.** Rev. Ibero-Amer. Ciênc. Ambient., v. 13, n. 7, 2023.

PADILHA, et al. **Análise da utilização de três diferentes coagulantes na remoção da turbidez de água de manancial de abastecimento.** Monografia/Unicesumar, 2011.

PINTO, José Maria; MIRANDA, J. H. de; PIRES, R. C. de M. **Qualidade da água.** Irrigação, Piracicaba: FUNEP, 2001.

PROBST, Philipp; WRIGHT, Marvin; BOULESTEIX, Anne-Laure. ***Hyperparameters and Tuning Strategies for Random Forest.*** arXiv preprint, 2018.

ROCHA, Gisele Verônica das Mercês. **Avaliação da quitosana como coagulante na redução da turbidez de efluentes minerários.** 2025. Monografia (Graduação em Engenharia Civil) – Universidade Federal de Ouro Preto, Escola de Minas, Departamento de Engenharia Civil, Ouro Preto, 2025.

RUSSELL, Stuart; NORVIG, Peter. ***Artificial Intelligence: A Modern Approach.*** 4. ed. Hoboken: Pearson, 2020.

SÁVIO, Luiz. **O aprendizado de máquina na perspectiva da automação industrial.** KPMG Business Insights, São Paulo, ed. 103, jul. 2023. Disponível em: https://home.kpmg/br/pt/home/insights.html. Acesso em: 1 ago. 2025.

SCHMIDT, Aline Ruth. **Análise da utilização do policloreto de alumínio (PAC) e sulfato de alumínio na eliminação de turbidez de água de abastecimento.** Trabalho de Conclusão de Curso (Especialização) – Universidade Tecnológica Federal do Paraná, Medianeira, 2014.

SILVA, C. A. A. et al. **Efeito da estação climática sobre a turbidez da água de um manancial: implicações para a gestão da qualidade da água.** Revista de Geografia, 2021.

SOUZA, A. P. C.; SOUZA, E. A. M.; PEREIRA, N. C. P. **Análise da utilização do coagulante policloreto de alumínio (PAC) na remoção da cor, turbidez e DQO de efluente de lavanderia têxtil.** Blucher Chemical Engineering Proceedings, v. 1, n. 2, 2015.

VEOLIA. **Descritivo de processo – ETA e desidratação de lodo.** [S.l.]: Veolia Water, 2012.