Universidade Federal de Ouro Preto
Escola de Minas
Engenharia de Controle e Automação

Júlia Silva Reis

Sensor Virtual para Classificação da Sílica no Produto de um Processo de Flotação de Minério de Ferro

Ouro Preto

Júlia Silva Reis

Sensor Virtual para Classificação da Sílica no Produto de um Processo de Flotação de Minério de Ferro

Monografia apresentada ao Curso de Engenharia de Controle e Automação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Engenheira de Controle e Automação.

Orientador: Me. Lucas Andery Reis
Coorientador: Dr. José Agnaldo da Rocha Reis

Ouro Preto
2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
DEPARTAMENTO DE ENGENHARIA CONTROLE E AUTOMACAO

FOLHA DE APROVAÇÃO

Júlia Silva Reis

Soft Sensor para Classificação da Sílica no Produto de um Processo de Flotação de Minério de Ferro

Monografia apresentada ao Curso de Engenharia de Controle e Automação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Engenheira de Controle e Automação

Aprovada em 02 de março de 2026

Membros da banca

Mestre - Lucas Andery Reis - Orientador (Vale SA)
Doutor - Agnaldo José da Rocha Reis - Coorientador (Universidade Federal de Ouro Preto)
Doutor - Paulo Marcos de Barros Monteiro - Examinador (Universidade Federal de Ouro Preto)
Doutor - Thomás Vargas Barsante Pinto - Examinador (Instituto Tecnológico Vale)

Lucas Andery Reis e Agnaldo José da Rocha Reis, orientadores do trabalho, aprovaram a versão final e autorizaram seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 07/03/2026.

Documento assinado eletronicamente por **Agnaldo Jose da Rocha Reis**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 07/03/2026, às 23:30, 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 **1070940** e o código CRC **957B9BA4**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591533 - www.ufop.br

Agradecimentos

Pela realização deste trabalho, agradeço, primeiramente, a Deus, que sempre esteve presente em minha vida, guiando meus passos, renovando minhas forças e me dando coragem para seguir em frente, mesmo nos momentos mais difíceis.

Aos meus pais, Maria José e Francisco, que são minha base, agradeço por todo o amor, apoio, dedicação e por terem segurado em minhas mãos nos momentos de incerteza, nunca me deixando desistir.

Ao Gustavo, por compartilhar comigo praticamente toda a trajetória do curso, pelos momentos vividos, pelo companheirismo constante, pelo apoio nos dias difíceis e pelo carinho.

À minha família, que sempre esteve ao meu lado: à minha avó, que está sempre torcendo por mim; à minha irmã, Ana Carolina, por todo o amor; à minha tia Daliandra, por sua presença, pelos conselhos e por todo o cuidado; aos meus padrinhos, Diana e Vagner, que sempre torcem por mim, me auxiliam e se alegram com minhas conquistas; à minha tia Rosa, pelo carinho e por ter sido essencial para possibilitar minha viagem acadêmica ao Peru; à minha prima Gabrielle, por ser uma presença motivadora e uma grande amiga; e às minhas tias Maria de Lourdes e Auxiliadora, e ao meu tio Genésio, por todas as orações e carinho.

À Creusa e ao Serginho, por serem amigos tão especiais da minha família e por todo o apoio, carinho e incentivo ao longo dessa jornada.

A todos os meus amigos, que estiveram presentes nos momentos bons e difíceis, oferecendo conselhos e tornando essa caminhada mais leve.

Aos meus orientadores, Lucas e Agnaldo, por todos os ensinamentos, pela paciência, pela disponibilidade e pela confiança em meu trabalho. Sua orientação foi fundamental para a concretização deste projeto.

À VALE S.A., aos professores e ao Departamento de Controle e Automação, por terem me proporcionado tantas oportunidades e aprendizados ao longo da minha trajetória como estudante de graduação.

Por fim, agradeço, de coração, a todos que, de alguma forma, fizeram parte dessa trajetória e contribuíram para a realização deste trabalho.

Resumo

A flotação reversa de minério de ferro visa reduzir o teor de sílica no concentrado. Porém, o controle eficaz é limitado pela disponibilidade descontínua dos resultados laboratoriais, obtidos apenas a cada duas horas. Propõe-se neste trabalho um sensor virtual baseado em aprendizado de máquina para classificar em tempo real o teor de sílica no concentrado da planta de flotação da Mina de Timbopeba da Vale S.A., utilizando dados históricos de processo. A metodologia seguiu etapas de inspeção, pré-processamento, análise estatística e modelagem. Após análise exploratória, selecionaram-se 21 variáveis como entradas do modelo. Para ser possível utilizar as variáveis de laboratório como entradas do modelo, aplicou-se uma estratégia de médias móveis. O modelo utilizado foi o *Random Forest*, avaliado em classificação binária, cujos rótulos escolhidos foram "Normal" ($\leq 2,5\%$), para porcentagens de sílica no concentrado dentro da faixa de especificação, e "Alta" ($> 2,5\%$), para porcentagens acima do especificado; e multiclasse, para os teores de sílica no concentrado considerados baixos ($< 1,5\%$), moderados (entre $1,5\%$ e $2,5\%$) ou altos ($> 2,5\%$). Os dados apresentaram desbalanceamento acentuado. Para mitigá-lo, utilizaram-se duas estratégias: ponderação por classe e balanceamento híbrido combinando *oversampling* sintético (SMOTE) com limpeza (Tomek Links). Conclui-se que o sensor virtual desenvolvido, em especial com balanceamento híbrido, é viável para fornecer estimativas categóricas em tempo real do teor de sílica. Assim, sua saída pode ser integrada a sistemas de controle avançado para ajustes ágeis na dosagem de reagentes e no nível das células, otimizando o processo, reduzindo custos e melhorando a qualidade do concentrado final.

Palavras-chave: Sensor virtual; Classificação; Flotação Reversa; Minério de Ferro; Sílica; Aprendizado de Máquina; *Random Forest*; SMOTE.

Abstract

Reverse iron ore flotation aims to reduce the silica content in the concentrate. However, effective control is limited by the discontinuous availability of laboratory results, which are obtained only every two hours. This study proposes a machine learning–based soft sensor to classify, in real time, the silica content in the concentrate of the flotation plant at the Timbopeba Mine operated by Vale S.A., using historical process data. The methodology comprised inspection, preprocessing, statistical analysis, and modeling stages. After exploratory analysis, 21 variables were selected as model inputs. In order to incorporate laboratory variables as model inputs, a moving average strategy was applied. The model adopted was the Random Forest algorithm, evaluated under both binary and multiclass classification approaches. In the binary setting, the selected labels were "Normal" ($\leq 2.5\%$), for silica percentages within the specification range, and "High" ($> 2.5\%$), for values above the specified limit. In the multiclass setting, silica levels in the concentrate were categorized as low ($< 1.5\%$), moderate (between 1.5% and 2.5%), or high ($> 2.5\%$). The dataset exhibited significant class imbalance. To mitigate this issue, two strategies were employed: class weighting and hybrid resampling combining synthetic oversampling (SMOTE) with data cleaning (Tomek Links). It can be concluded that the developed soft sensor, particularly when combined with hybrid resampling, is feasible for providing real-time categorical estimates of silica content. Therefore, its output can be integrated into advanced control systems to enable rapid adjustments in reagent dosage and cell level, optimizing the process, reducing costs, and improving the quality of the final concentrate.

Key-words: Soft sensor; Classification; Reverse Flotation; Iron Ore; Silica; Machine Learning; Random Forest; SMOTE.

Lista de figuras

Lista de abreviaturas e siglas

SP	Setpoint
RGB	Red, Green, Blue
VP	Verdadeiro Positivo
FP	Falso Positivo
VN	Verdadeiro Negativo
FN	Falso Negativo

Sumário

1 Introdução

Em usinas de beneficiamento de minérios, a flotação é necessária para a produção de concentrados minerais mais valiosos, por meio da separação dos minerais de ganga dos minerais de interesse. Existem diversos processos de flotação, sendo o mais comum o processo de flotação por espumas (*froth flotation*) (CHAVES; LEAL FILHO; BRAGA, 2018). Nesse processo, dois métodos, direto ou reverso, podem ser utilizados para separar o mineral de interesse.

O método da flotação direta consiste, segundo Chaves, Leal Filho e Braga (2018), em flotar os minerais de interesse, que são separados na espuma, enquanto os minerais de ganga permanecem na polpa. No método da flotação reversa ocorre o contrário, uma vez que os minerais de interesse acompanham o fluxo da polpa e os minerais de ganga são flotados. No entanto, para que essa separação ocorra com uma boa recuperação do mineral de interesse, o uso de reagentes é essencial no processo.

Os principais reagentes utilizados em um processo de flotação são os coletores e os depressores. O objetivo da utilização do coletor é transformar a superfície de um mineral em hidrofóbica, de forma que ele possa aderir-se às bolhas do processo e flotar. No entanto, para que o coletor escolha uma espécie mineral sem causar modificações nas outras, pode ser utilizado o reagente depressor, que deprime a ação de coleta em partículas indesejadas (CHAVES; LEAL FILHO; BRAGA, 2018). Além dos coletores e depressores, podem ser utilizados reagentes ativadores, reguladores de pH, dispersantes e espumantes.

De acordo com Luz, Sampaio e França (2010), a flotação ocorre em equipamentos denominados células (máquinas) ou colunas de flotação, que se diferenciam, principalmente, pela geometria e pelo funcionamento. Isso porque, nas células, é utilizado um dispositivo móvel, chamado máquina de flotação, que produz bolhas de ar enquanto é agitado. Por outro lado, as colunas possuem em sua base um mecanismo de aeração, que gera as bolhas ascendentes, responsáveis por flotar o material desejado.

Para processos industriais, uma única etapa de flotação não é suficiente para conseguir a recuperação de minério almejada (LUZ; SAMPAIO; FRANÇA, 2010). Por esse motivo, circuitos utilizando um conjunto de equipamentos de flotação são projetados, sendo que estes possuem, normalmente, três estágios: *rougher*, *cleaner* e *scavenger*. Nesse sentido, cada uma das fases influencia na produção do valor desejado de concentrado.

Os circuitos de flotação são conhecidos por serem não lineares, multivariáveis e por possuírem parâmetros mutáveis (LIAO et al., 2011). Sendo assim, diante da complexidade do processo, a implementação de estratégias de controle eficazes nas plantas de flotação torna-se essencial. Nesse sentido, a escolha de um modelo adequado de controle é fundamental para alcançar um desempenho estável e com melhores resultados durante a operação das plantas.

Dentre os tipos de controles utilizados estão os regulatórios, que trabalham com um

ponto de operação fixo, conhecido como *setpoint* (SP), que busca ser mantido mesmo na presença de perturbações (CAMPOS, 2010). Existem também os controles avançados, que calculam valores variáveis de SP (CAMPOS; GOMES; PEREZ, 2013), de acordo com as condições de campo.

Na planta objeto deste estudo, os controles avançados existentes são desenvolvidos utilizando a técnica de lógica *fuzzy* para a dosagem de amido e amina, e para a determinação do SP de nível das células de flotação. Dentre outras variáveis de processo, esses controles avançados utilizam o percentual de sílica no concentrado como entrada, que é obtida por meio de análises laboratoriais, em intervalos de 2 horas.

Este trabalho concentra-se na camada de controle avançado da pirâmide de automação, baseada em Strutzel (2014), apresentada na Figura 1. Observa-se que o controle regulatório opera com base em dados, coletados por sensores instalados em campo com o objetivo de manter as variáveis controladas nos *setpoints* pré-definidos. Por sua vez, o controle avançado calcula valores de SP para serem utilizados pelo controle regulatório. Para isso, além dos dados fornecidos por sensores, o controle avançado necessita de informações sobre os resultados da operação, coletados por meio de análises laboratoriais. Desse modo, eles são capazes de determinar o ponto ótimo de operação, possibilitando ajustes dinâmicos que otimizam o desempenho do processo. Entretanto, análises laboratoriais do material processado na flotação ocorrem em intervalos de horas, o que impacta na tomada de ação de curto prazo pelo sistema de controle.

Figura 1 – Pirâmide de Controle de Processos. Fonte: Adaptada de Strutzel (2014).

Neste trabalho, tem-se como objetivo desenvolver um sensor virtual capaz de classificar resultados de sílica no concentrado em tempo real. Tal sensor virtual receberá como entradas a vazão de alimentação de polpa, as dosagens de amido e amina, os níveis das células de flotação, entre outras variáveis laboratoriais, e produzirá em sua saída a classificação em faixas do percentual de sílica no concentrado. Valida-se o algoritmo proposto a partir de um estudo de caso com dados reais de uma planta de flotação da usina de beneficiamento de Timbopeba da Vale S.A., localizada em Ouro Preto-MG.

1.1 Justificativas e Relevância

Em um processo de flotação, amostras são retiradas, periodicamente, e enviadas a um laboratório, que realiza a análise, por exemplo, do teor de sílica no concentrado. Após a análise, as variáveis do processo, como reagentes, nível da célula/coluna de flotação, velocidade da flotação, entre outras, são ajustadas para manter o processo em um valor desejado de concentração. No entanto, um período de duas horas, no caso em estudo, é um tempo considerável, fazendo com que os operadores avaliem as condições de trabalho por meio da observação das características da espuma. Essa demora, segundo Zhang e Gao (2021), indica que o tempo gasto para análise das amostras pode causar detecção tardia de eventos não desejados, como o aumento do percentual de ferro no rejeito e de sílica no concentrado, que acarretam em perdas econômicas e de qualidade do produto.

Na etapa de flotação da usina de beneficiamento de minério de ferro estudada, os controles avançados existentes para o ajuste de nível das células fazem uma comparação entre o valor real de sílica no concentrado, disponibilizado de 2 em 2 horas, e o SP para essa variável. Conforme o resultado de erro, os controles avançados atuam ajustando o ponto de operação do processo. No entanto, indicadores de utilização mostram baixa utilização dos sistemas de controles avançados pelos colaboradores da operação, que desabilitam-os justificando perda de qualidade no produto da flotação. Dessa forma, a equipe opera, na maior parte do tempo, utilizando apenas o controle regulatório do processo, onde são capazes de ajustar manualmente o SP do processo, conforme a necessidade.

A relevância deste trabalho encontra-se na busca de uma solução para contornar a lacuna de tempo entre a coleta de amostras e a entrega dos resultados de laboratório referentes ao percentual de sílica no concentrado, uma vez que este teor mineral é um indicador-chave de desempenho. Além disso, este estudo contribui para o avanço tecnológico ao propor a utilização de um sensor virtual para classificar, em tempo real, os resultados do processo. Assim, a saída do sensor virtual pode ser utilizada como variável de entrada do controlador avançado para aumentar sua performance. Consoante ao Ai et al. (2021), pequenos aumentos de desempenho em um processo de flotação podem gerar elevados benefícios econômicos.

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver um sensor virtual para classificar os teores de sílica no concentrado da planta de flotação de minério de ferro estudada, em tempo real, com base em dados históricos de sensores e de análises laboratoriais.

1.2.2 Objetivos Específicos

- Estudar a dinâmica do processo de flotação reversa de minério de ferro e investigar as variáveis envolvidas;
- Coletar dados históricos dos sensores e das análises laboratoriais da planta de Timbopeba para a modelagem do sensor virtual;
- Realizar o pré-processamento da base de dados e verificar sua adequação para a modelagem do sensor virtual;
- Selecionar técnicas de aprendizado de máquina adequadas para o desenvolvimento do sensor virtual e avaliar seus desempenhos.

1.3 Organização e estrutura

Este trabalho está organizado da seguinte forma: no Capítulo 2 é apresentada a fundamentação teórica, detalhando o circuito de flotação, a teoria por trás da estratégia proposta e abordando estudos relacionados. O Capítulo 3 descreve os materiais e métodos adotados, incluindo as estratégias de pré-processamento dos dados e a metodologia empregada para a seleção, o treinamento e a validação dos modelos. No Capítulo 4, são apresentados e discutidos os resultados obtidos para os modelos propostos. Por fim, o Capítulo 5 reúne as conclusões e as perspectivas para trabalhos futuros.

2 Fundamentação Teórica

2.1 Processo de Flotação

Nesta seção são apresentados os conceitos relacionados ao processo de flotação em células. Além disso, é apresentado o detalhamento do circuito de flotação da usina Timbopeba, pertence à Vale S.A..

2.1.1 Células de Flotação

As células de flotação são equipamentos projetados para tratar a polpa de minério, separando o mineral de interesse do rejeito. Para isso, a célula recebe a polpa mineral em uma de suas faces laterais e descarrega-a na face oposta (CHAVES; LEAL FILHO; BRAGA, 2018). O compartimento de descarga é responsável por realizar o transbordo do material flotado por calhas coletoras e o escoamento do material não flotado pelo fundo da célula.

Apresenta-se, na Figura 2, uma máquina de flotação, sendo ela um dispositivo que consiste de um rotor instalado dentro da célula de flotação (CHAVES; LEAL FILHO; BRAGA, 2018). A função da máquina de flotação é manter a polpa suspensa, por meio da movimentação do rotor. Essa movimentação produz uma região de pressão negativa que, em muitos casos, é suficiente para prover o ar necessário para o processo. Esse ar é responsável pelo carregamento das partículas a serem flotadas. Para isso, ele é fragmentado em pequenas bolhas por uma peça instalada ao redor do rotor, denominada estator (CHAVES; LEAL FILHO; BRAGA, 2018).

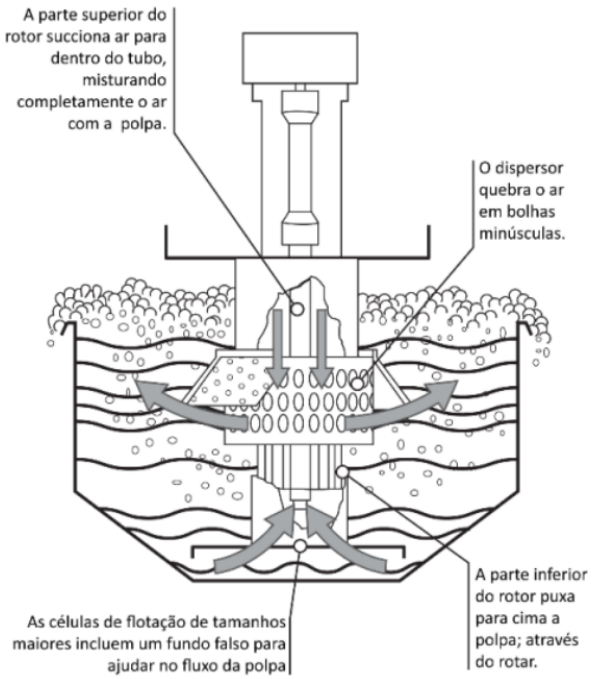


Figura 2 – Esquemático de uma máquina de flotação. Fonte: Chaves, Leal Filho e Braga (2018).

Ainda, segundo Chaves, Leal Filho e Braga (2018), a flotação pode ocorrer em apenas uma célula; no entanto, é comum a utilização de um conjunto de células para aumentar a recuperação mineral. Sendo assim, o processo é dividido em três estágios principais: *rougher*, *cleaner* e *scavenger*.

Os três estágios são descritos por Chaves, Leal Filho e Braga (2018). Em primeiro momento, a polpa passa pela etapa *rougher*, que realiza a separação inicial, resultando em um concentrado pobre e em um rejeito que ainda contém boa parte do mineral desejado. A etapa *cleaner* executa a limpeza do concentrado proveniente da etapa *rougher*, produzindo um concentrado rico e um rejeito de alto teor. Por outro lado, a etapa *scavenger* processa o rejeito *rougher*, originando um rejeito final, que é descartado, e um concentrado contendo traços do mineral desejado. Tendo em vista que o processo de flotação é cíclico, existe uma carga circulante, formada pelo rejeito *cleaner* e pelo concentrado *scavenger*, que retorna para o estágio *rougher* e é reprocessada. Em alguns casos, torna-se necessária a implementação de mais etapas de limpeza, chamadas *recleaner*.

2.1.2 Circuito de Flotação Reversa de Minério de Ferro da VALE - Mina de Timbopeba

Na Figura 3, ilustra-se o fluxograma do processo de flotação da usina de beneficiamento de minério de ferro da Vale S.A., na mina de Timbopeba (VALE, 2021). Trata-se de um circuito composto por duas linhas paralelas, onde cada uma possui os estágios *rougher*, *cleaner*, *recleaner* e *scavenger*.

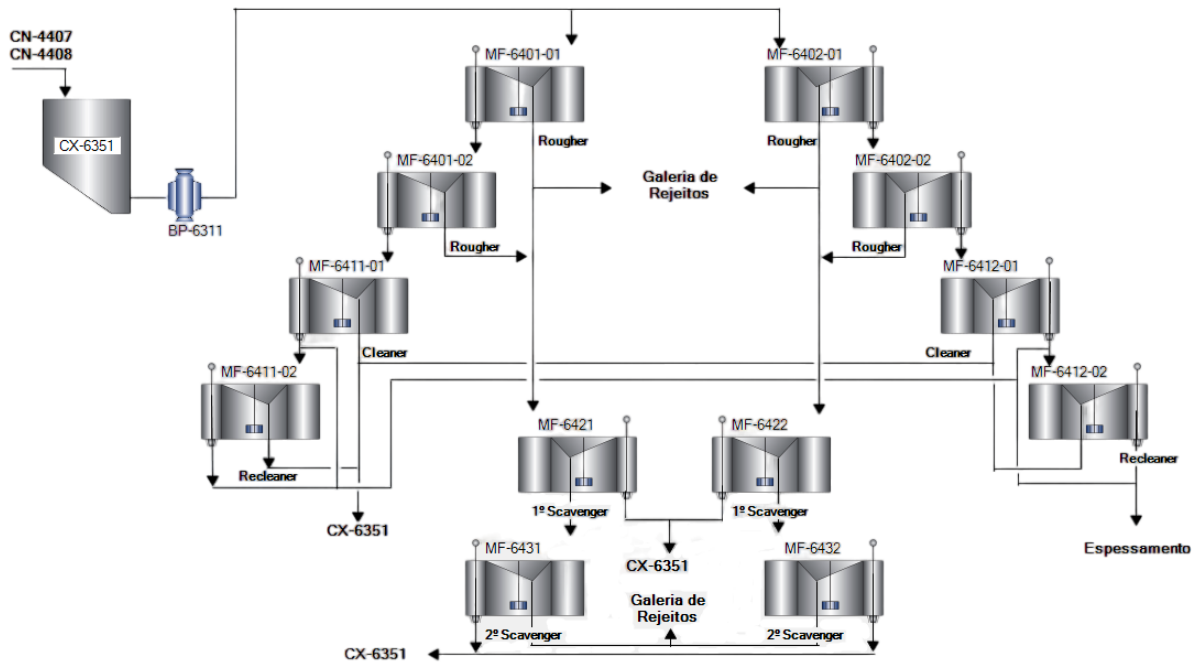


Figura 3 – Fluxograma do processo de flotação reversa de minério de ferro da usina de Timbopeba. Fonte: VALE (2021)

Cada linha do circuito é formada por seis células de flotação, sendo duas células *rougher*, uma célula *cleaner*, uma célula *recleaner* e duas células *scavenger*. A alimentação do circuito é proveniente do produto do processo de condicionamento. Nesse sentido, observa-se que o material é bombeado na primeira célula *rougher*. No circuito, os materiais concentrados seguem o caminho: primeira *rougher*, segunda *rougher*, *cleaner* e *recleaner*. O concentrado final, da célula *recleaner*, segue para o processo seguinte de beneficiamento do minério de ferro, o espessamento.

Por outro lado, os rejeitos de ambas as células *rougher* são designados para a primeira *scavenger*. Em seguida, o rejeito proveniente da primeira *scavenger* é direcionado para a segunda *scavenger*, que produz o rejeito final, direcionado para a cava de rejeitos.

O circuito também possui uma carga circulante, uma vez que os rejeitos produzidos nas células *cleaner* e *recleaner*, assim como os concentrados da primeira e da segunda *scavenger*, são direcionados para a caixa proveniente do processo de condicionamento, para serem concentrados novamente pelo circuito de flotação.

No processo são adicionados reagentes depressores e coletores. O reagente depressor utilizado é o amido, que é acrescentado ao beneficiamento nas etapas de condicionamento, sendo, também, comumente adicionado na etapa *scavenger*, para melhorar a seletividade das partículas. Por outro lado, a amina, coletor utilizado, é adicionada, normalmente, nas calhas de alimentação dos estágios *rougher* e *cleaner*.

Amostras do produto do circuito de flotação são coletadas periodicamente e são enviadas para análises laboratoriais. Os resultados do laboratório químico são registrados no PIMS, de 2 em 2 horas. As análises entregam os resultados do teor de sílica no concentrado e do teor de ferro no rejeito.

Os teores de sílica no concentrado e de ferro no rejeito são alguns dos indicadores utilizados que demonstram se a qualidade do produto da flotação está com uma qualidade ruim ou boa. Nesse sentido, quando os teores estão acima do especificado, o resultado é ruim. Sendo assim, esses indicadores são utilizados por alguns controladores avançados como variáveis de observação, como por exemplo pelo controle da dosagem de reagentes e pelo controle de nível de polpa.

A dosagem dos reagentes (em g/t) é ajustada com base nos teores de sílica no concentrado e de ferro no rejeito. Para o depressor, o ajuste relaciona-se às variações do teor de ferro no rejeito. Nesse caso, quando o teor está acima do especificado, mais amido é adicionado. Por outro lado, o ajuste da quantidade de amina adicionada é baseado no resultado do teor de sílica no concentrado. Assim, quando esse teor está acima do especificado, mais amina é adicionada.

Conforme Figura 4, o processo do controle de nível das células de flotação inicia-se na leitura do percentual de sílica do concentrado, no qual seu erro em relação ao *setpoint* é calculado. O erro negativo indica que o teor de sílica no concentrado está acima do especificado. Nesse caso, é necessário aumentar a velocidade de transbordo da célula, cujo valor é coletado em tempo real por câmeras instaladas nas células. Para isso, o controlador *fuzzy* incrementa o

setpoint de nível da célula para eliminar o excesso de sílica.

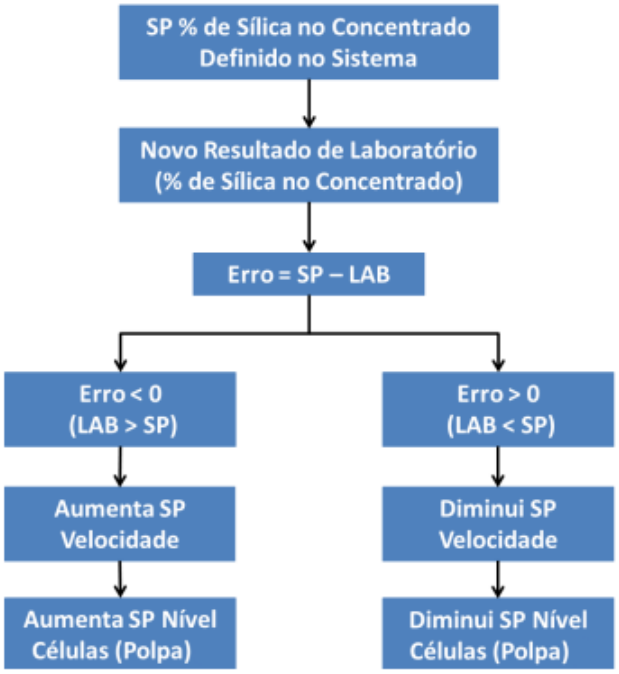


Figura 4 – Regras do controle *fuzzy* de nível das células de flotação. Fonte: VALE (2021)

Além dos resultados da flotação, são armazenadas no PIMS todas as variáveis presentes no processo, como os estados dos equipamentos, das malhas de controle e dos parâmetros do processo. Essa armazenagem cria dados históricos que possibilitam avaliações operacionais e de engenharia sobre a eficiência do processo.

2.2 Sensores Virtuais

Em um processo industrial, existem variáveis difíceis de serem medidas por sensores físicos, instalados em campo. Dessa forma, a aferição de valor dessas variáveis é comumente realizada por meio da análise de amostras retiradas do processo. No entanto, segundo Lipták e Venczel (2016), como a análise laboratorial é infrequente e demorada, certas desvantagens são verificadas, como o aumento do custo de produção, produtos fora das especificações e desperdício de recursos.

Uma alternativa à realização de ensaios laboratoriais para mensurar os valores de variáveis de difícil medição é a utilização de sensores virtuais, também conhecidos como instrumentos virtuais, sensores inferenciais, sensores inteligentes ou sensores baseados em *software* (DINIZ, 2020). De acordo com Lipták e Venczel (2016), os sensores virtuais são aplicados em situações que incluem longos atrasos de medição, custo elevado do dispositivo ou técnica de medição, baixa confiabilidade e falta ou indisponibilidade de dispositivos adequados para a medição.

Em geral, a implementação de sensores virtuais pode ser orientada por modelo ou por dados (KADLEC; GABRYS; STRANDT, 2009). Na orientação por modelo, eles descrevem o contexto físico e químico do processo, exigindo conhecimento profundo do processo para deduzir

suas equações (LIPTÁK; VENCZEL, 2016). Por outro lado, sensores inteligentes baseados em dados utilizam dados historiados das plantas, sendo comumente mais utilizados, uma vez que, usualmente, descrevem melhor as condições reais dos processos.

No presente trabalho, busca-se desenvolver um sensor virtual baseado em dados, com o intuito de classificar o teor de sílica no concentrado da flotação de minério de ferro. Para tanto, as entradas serão a vazão de alimentação de polpa, as dosagens de amido e amina, os níveis das células de flotação, entre outras, assim como variáveis advindas de análises laboratoriais; e a saída será a categorização da sílica no concentrado, em classes como baixa, moderada ou alta.

2.2.1 Metodolodia para o desenvolvimento de sensores virtuais

Nesta subseção, apresenta-se a metodologia proposta por Kadlec, Gabrys e Strandt (2009) para o desenvolvimento de sensores virtuais. A metodologia, apresentada na Figura 5, é composta pelas etapas: primeira inspeção dos dados, seleção de dados históricos e identificação de estados estacionários, pré-processamento dos dados, seleção, treinamento e validação do modelo e, por fim, manutenção do sensor virtual.

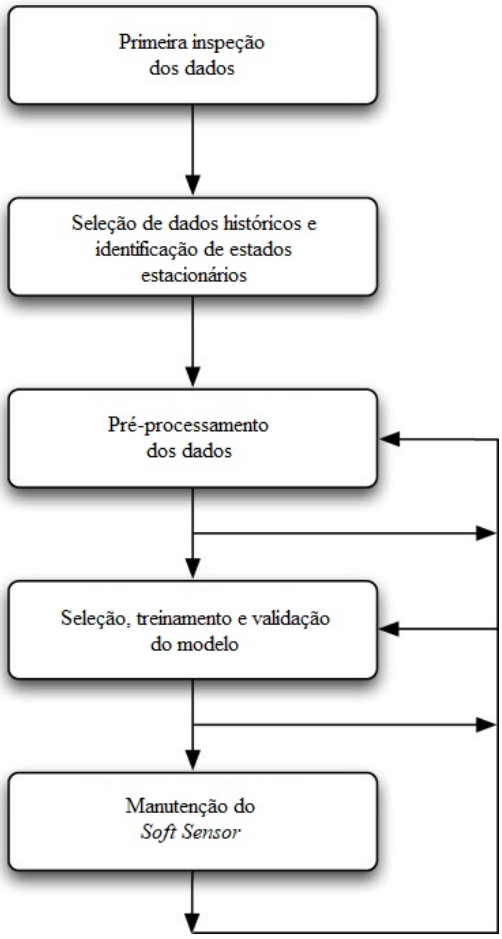


Figura 5 – Metodologia de desenvolvimento de sensores virtuais. Fonte: Adaptado de Kadlec, Gabrys e Strandt (2009).

Primeira inspeção dos dados

A primeira etapa tem como objetivo inspecionar os dados, de forma a tratar possíveis problemas entre eles. Segundo Diniz (2020), esses tratamentos envolvem a detecção do congelamento de variáveis ou de variáveis cujos instrumentos de medição estejam inativos. Dessa forma, esses dados, cujos comportamentos não contribuem para a previsão da variável de interesse, podem ser desconsiderados. Além disso, nesta fase, deve-se analisar cuidadosamente a variável alvo com o intuito de verificar se sua modelagem é possível (KADLEC; GABRYS; STRANDT, 2009).

Seleção de dados históricos e identificação de estados estacionários

Nesta etapa, os dados históricos que passaram pela inspeção inicial são selecionados, com o intuito de serem utilizados para o treinamento e para a avaliação do modelo (KADLEC; GABRYS; STRANDT, 2009). Além disso, são identificados os estados estacionários das variáveis, sendo que, na maioria das vezes, eles são selecionados para uma modelagem adicional.

Pré-processamento dos dados

Nesta parte da metodologia são aplicadas técnicas de pré-processamento nos dados selecionados, para que eles possam ser melhor processados pelo modelo. Em processos industriais, os dados geralmente passam por diversas etapas de pré-processamento, como ilustra o fluxo da Figura 5.

De acordo com Kadlec, Gabrys e Strandt (2009), as etapas de pré-processamento incluem: normalização dos dados para média zero e variância unitária, tratamento de dados ausentes, detecção, seleção de variáveis relevantes, substituição de *outliers* que, segundo Diniz (2020), são dados que se desviam da distribuição normal em uma mesma amostra, entre outras.

Esta etapa requer atenção especial e, normalmente, é repetida diversas vezes até que o desenvolvedor avalie a adequação dos dados para serem utilizados no treinamento e na validação do modelo (KADLEC; GABRYS; STRANDT, 2009).

Seleção, treinamento e validação do modelo

Selecionar um bom modelo é essencial para o desempenho do sensor inteligente. Isso porque, como mencionado por Kadlec, Gabrys e Strandt (2009), o modelo é o motor do sensor virtual, interferindo diretamente em seu desempenho. No entanto, não há uma abordagem consolidada para a escolha do melhor modelo de desenvolvimento. Por esse motivo, os desenvolvedores identificam o melhor modelo com base em suas experiências e em testes.

Lipták e Venczel (2016) descrevem que uma alternativa é iniciar avaliando a escolha entre modelos estáticos ou dinâmicos. Em processos industriais, modelos dinâmicos tendem a descrever melhor as características não lineares das plantas. Isso porque, conforme Lipták e Venczel (2016), mesmo em estados estacionários os comportamentos dinâmicos ocorrem nos

processos, sendo observados durante estágios de parada, partida e despressurização. Além disso, Kadlec, Gabrys e Strandt (2009) sugerem iniciar desenvolvendo modelos lineares, que são mais simples. Após, a implementação de modelos não lineares, que são mais complexos, podem ser utilizados quando nota-se possibilidade de melhoria de desempenho.

Uma vez que a técnica para o desenvolvimento do sensor virtual tenha sido definida, realiza-se, em seguida, o treinamento do modelo, utilizando uma parte dos dados selecionados. Nesse caso, para verificar se o modelo está tendo boa generalização, ou seja, se está conseguindo ter bons resultados com dados desconhecidos, utiliza-se outra parte dos dados selecionados, conhecida como conjunto de teste. Segundo Lipták e Venczel (2016), quando o modelo possui um bom desempenho no treinamento e um desempenho ruim no teste, esse comportamento é denominado *overfitting*, ou sobreajuste.

Para evitar que problemas como o *overfitting* ocorram em um processo real, validações do modelo são implementadas, utilizando o restante dos dados selecionados, que não foram usados nas etapas de treino e teste. Para isso, podem ser utilizados métodos quantitativos, como o Erro Quadrático Médio (do termo em inglês *Mean Squared Error*, MSE), que calcula a média da distância quadrática entre o valor aferido e o valor real (KADLEC; GABRYS; STRANDT, 2009). O MSE é dado pela Equação 2.1, em que y_i é o valor real, $\hat{y}_i$ é o valor previsto e n é o número de amostras.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.1}$$

Ainda que a etapa de validação seja muito importante, ela é frequentemente subestimada (LIPTÁK; VENCZEL, 2016). Além disso, Diniz (2020) afirma que, apesar de subjetiva, a avaliação adicional de especialistas quanto ao desempenho do modelo de sensor virtual desenvolvido também é importante.

Manutenção do sensor virtual

Tendo em vista que a dinâmica de processos industriais é variável, dados de um período anterior do processo podem não ser eficientes a longo prazo. Dessa forma, torna-se necessário realizar a manutenção do sensor virtual regularmente, para evitar sua deterioração (KADLEC; GABRYS; STRANDT, 2009). As manutenções podem ser realizadas por meio da adaptação do modelo, alterando os dados do processo por outros mais atuais, ou por meio do desenvolvimento de um novo modelo.

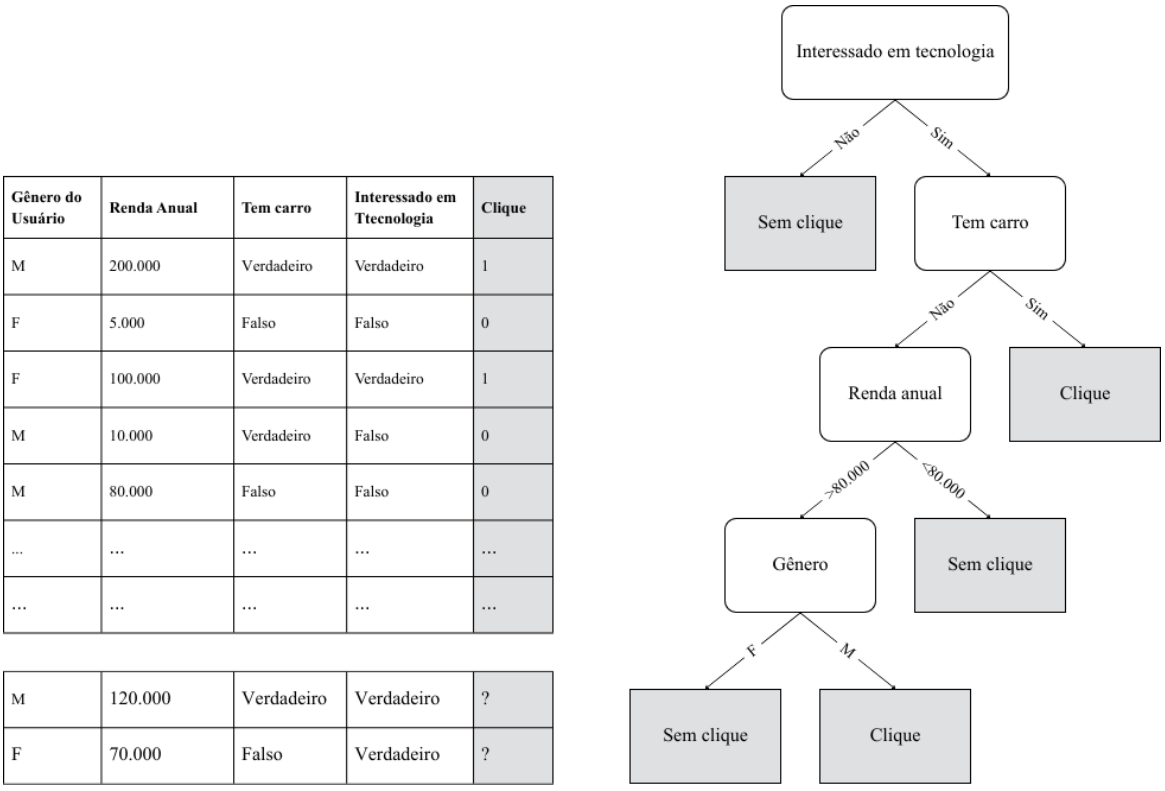
2.2.2 Random Forest

Segundo Liu (2024), floresta aleatória (*random forest*) é uma técnica baseada na agregação bootstrap (*bootstrap aggregating*). Sendo assim, uma *random forest* funciona combinando várias árvores de decisão. Em vez de treinar apenas uma árvore, o método cria muitas árvores diferentes,

cada uma construída a partir de amostras escolhidas de forma aleatória dos dados originais. Após o treinamento de todas as árvore, suas respostas são reunidas. Para problemas de classificação, o resultado é decidido pela maioria das árvores.

Tal processo auxilia na redução de problemas enfrentados por uma única árvore de decisão, como a tendência de se ajustar demais aos dados de treinamento (*overfitting*). Sendo assim, com a junção de várias árvores o modelo fica mais estável e costuma desempenhar melhor. No entanto, o ganho é pequeno caso todas as árvores sejam muito parecidas. Visando contornar esse problema, a *random forest* faz com que cada árvore considere apenas uma parte das variáveis disponíveis em cada divisão, de forma que elas sejam diferentes entre si, aumentando a qualidade das previsões.

A visualização de uma única árvore é fundamental para compreender a motivação do *random forest*. Na Figura 6, ilustra-se como uma árvore de decisão funciona na prática. Esse exemplo apresenta um processo de classificação de usuários em relação ao clique em um anúncio, o que destaca como condições sucessivas levam a um resultado final. Assim, ao combinar diversas árvores como essa, treinadas com diferentes subconjuntos de dados e atributos, obtém-se um modelo mais robusto e generalizável, o *random forest*.



Gênero do Usuário	Renda Anual	Tem carro	Interessado em Ttecnologia	Clique
M	200.000	Verdadeiro	Verdadeiro	1
F	5.000	Falso	Falso	0
F	100.000	Verdadeiro	Verdadeiro	1
M	10.000	Verdadeiro	Falso	0
M	80.000	Falso	Falso	0
...	...	...	...	...
...	...	...	...	...

M	120.000	Verdadeiro	Verdadeiro	?
F	70.000	Falso	Verdadeiro	?

Figura 6 – Árvore de decisão para prever clique/não clique em anúncios. Fonte: Adaptado de Liu (2024).

Em uma floresta aleatória, alguns hiperparâmetros críticos devem ser ajustados para obter melhor desempenho. São eles:

- **max_depth**: profundidade máxima das árvores individuais. Árvores muito profundas ten-

dem ao *overfitting*, enquanto árvores muito rasas podem levar ao subajuste (*underfitting*).

- **min_samples_split**: número mínimo de amostras necessário para realizar uma nova divisão em um nó. Valores muito pequenos favorecem o *overfitting*, enquanto valores muito grandes podem introduzir *underfitting*.
- **max_features**: número de características consideradas em cada busca pelo melhor ponto de divisão. Uma escolha comum é utilizar $\sqrt{m}$, onde m é o número total de atributos, ou alternativas como $\log_2(m)$.
- **n_estimators**: número de árvores na floresta. Em geral, quanto maior o número de árvores, melhor o desempenho, embora o custo computacional também aumente.

O *random forest* combina simplicidade, diversidade e robustez, sendo amplamente utilizado em problemas de classificação.

Modelos de Classificação

Segundo Liu (2024), a classificação é uma das principais abordagens do aprendizado supervisionado, cujo objetivo é desenvolver um modelo capaz de aprender uma regra que relacione corretamente as variáveis de entrada (preditoras ou características) com as de saída (rótulos). Esse modelo é formado por duas etapas, sendo a primeira de treinamento e a segunda de predição. Após o treinamento, o modelo pode ser utilizado para prever a classe de novas amostras, com base nas novas entradas. Na Figura 7, apresentam-se as etapas de treinamento e de predição de um modelo de classificação.

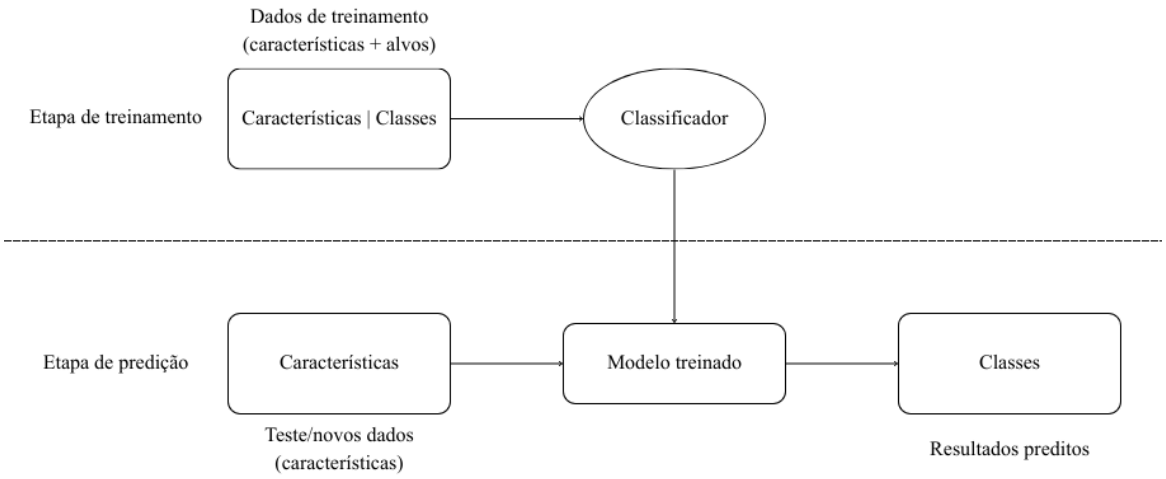


Figura 7 – Etapas de treinamento e predição em modelo de classificação. Fonte: Adaptado de Liu (2024).

Na etapa de treinamento, o classificador aprende a partir de dados rotulados com suas respectivas classes (alvos), gerando um modelo treinado. Na etapa seguinte, de predição, o modelo treinado recebe novas características e classifica-as.

De acordo Liu (2024), existem três tipos principais de classificação: binária, multiclasse e multi-rótulo. Na classificação binária, existem duas classes possíveis e cada amostra é atribuída a uma delas. Por outro lado, na classificação multiclasse existem três ou mais classes e cada amostra é associada a uma delas. Por fim, na classificação multi-rótulo, uma mesma amostra pode pertencer simultaneamente a múltiplas classes, sendo útil, por exemplo, na classificação de filmes por gênero.

Para estimar o desempenho de um classificador, a medida mais utilizada é a acurácia (A_{cc}), também conhecida como taxa de acerto (ESCOVEDO; KOSHIYAMA, 2020). Dessa forma, a acurácia representa a fração de exemplos corretamente classificados pelo modelo, sendo uma métrica simples e direta para avaliação de desempenho, dada pela Equação 2.2:

$$Acc(h) = 1 - Err(h), \tag{2.2}$$

em que $Acc(h)$ é a acurácia do classificador h e $Err(h)$ é a taxa de erro do classificador h, ou seja, a proporção de exemplos classificados incorretamente. Essa taxa é calculada conforme a Equação 2.3, em que n representa o número total de exemplos, y_i é o rótulo verdadeiro do i-ésimo exemplo e $h(i)$ é o rótulo previsto pelo classificador. O operador $\|y_i \neq h(i)\|$ retorna 1 se a expressão lógica for verdadeira e, caso contrário, retorna 0, de forma a contabilizar os erros de classificação.

$$Err(h) = \frac{1}{n}\sum_{i=1}^{n} \|y_i \neq h(i)\| \tag{2.3}$$

Métricas de Desempenho em Classificação

Embora a acurácia seja uma métrica amplamente utilizada, sua interpretação pode ser enganosa em cenários com distribuições de classe desbalanceadas. Nesses casos, métricas complementares tornam-se essenciais para uma avaliação mais criteriosa do desempenho do classificador. Dentre essas, destacam-se a precisão (do inglês *precision*), a revocação (do inglês *recall*) e a pontuação F_1 (do inglês *F1-Score*) (GÉRON, 2022).

Para compreender tais métricas, conceituar uma matriz de confusão (do inglês *confusion matrix*) torna-se fundamental. Nesse sentido, tem-se que uma matriz de confusão é uma tabela que contabiliza as previsões corretas e incorretas de um classificador, de acordo com a classe real e a classe prevista. Assim, cada linha da matriz representa uma classe real, enquanto cada coluna representa uma classe prevista. A partir dela, para uma classificação binária, podem-se extrair os dados referentes aos valores:

- **Verdadeiros Positivos (VP)**: Instâncias da classe positiva corretamente classificadas.
- **Falsos Positivos (FP)**: Instâncias da classe negativa incorretamente classificadas como positivas.

- **Falsos Negativos (FN)**: Instâncias da classe positiva incorretamente classificadas como negativas.
- **Verdadeiros Negativos (VN)**: Instâncias da classe negativa corretamente classificadas.

A partir desses valores, define-se precisão como a proporção de instâncias classificadas como positivas que realmente são positivas, conforme a Equação 2.4 (GÉRON, 2022):

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2.4}$$

Por outro lado, define-se revocação (ou sensibilidade) como a proporção de instâncias positivas reais que foram corretamente identificadas pelo classificador. Sua definição é dada pela Equação 2.5 (GÉRON, 2022):

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{2.5}$$

Em geral, existe uma relação conflitante entre precisão e revocação. Isso porque aumentar o limiar de decisão de um classificador geralmente eleva a precisão, uma vez que o modelo torna-se mais conservador ao prever a classe positiva, mas reduz a revocação, pois o modelo deixa de identificar mais instâncias positivas. De modo oposto, diminuir o limiar tende a aumentar a revocação e a reduzir a precisão (GÉRON, 2022).

Para combinar precisão e revocação em uma única métrica, utiliza-se o **F1-Score**, que corresponde à média harmônica das duas métricas. A média harmônica atribui maior peso a valores baixos, de modo que um alto valor de F_1 é apenas alcançado quando tanto a precisão quanto a revocação forem elevadas. A fórmula do F_1-Score é apresentada na Equação 2.6 (GÉRON, 2022):

$$F_1 = \frac{2}{\frac{1}{\text{Precisão}} + \frac{1}{\text{Revocação}}} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{2.6}$$

A escolha da métrica mais relevante depende do contexto do problema. Em aplicações onde os falsos positivos são muito custosos, prioriza-se a precisão. Já em situações onde os falsos negativos são críticos, a revocação é a métrica mais importante (GÉRON, 2022).

Métodos de Pré-processamento: Abordagens em Nível de Dados

Quando o conjunto de dados de treinamento apresenta distribuição desbalanceada entre as classes, os algoritmos tradicionais de classificação tendem a apresentar viés em favor da classe majoritária, resultando em maior taxa de erro de classificação para a classe minoritária (FERNÁNDEZ et al., 2018). Para mitigar esse problema, técnicas de pré-processamento em nível de dados, baseadas em métodos de amostragem (*sampling*), são amplamente utilizadas. Essas técnicas modificam a distribuição original dos dados, buscando equilibrar as classes ou

produzir uma distribuição mais adequada para as etapas subsequentes de aprendizado. Conforme destacado por Fernández et al. (2018), o rebalanceamento do conjunto de dados frequentemente leva a uma melhoria significativa no desempenho geral da classificação.

As técnicas de reamostragem podem ser categorizadas em três grupos principais (FERNÁNDEZ et al., 2018):

- ***Undersampling***: Cria um subconjunto do conjunto de dados original por meio da eliminação de instâncias, geralmente da classe majoritária.
- ***Oversampling***: Cria um superconjunto do conjunto de dados original por meio da replicação ou geração de novas instâncias da classe minoritária.
- **Métodos Híbridos**: Combinam abordagens de *undersampling* e *oversampling*.

Dentre os métodos de *oversampling*, a Técnica de Sobreamostragem de Minoria Sintética (do inglês *Synthetic Minority Over-sampling Technique* – SMOTE) destaca-se como uma das mais conhecidas e utilizadas (FERNÁNDEZ et al., 2018). Diferentemente da sobreamostragem aleatória, que apenas replica instâncias existentes, o SMOTE gera novas instâncias sintéticas para a classe minoritária por meio de interpolação entre exemplos positivos que estão próximos no espaço de características. Esse processo é ilustrado na Figura 8, em que apresenta-se a geração de pontos sintéticos no algoritmo SMOTE. Nesse caso, uma instância x_i da classe minoritária é selecionada, e seus k vizinhos mais próximos da mesma classe são identificados (pontos x_{i1} a x_{i4}). Assim, novas instâncias sintéticas (r_1 a r_4) são criadas ao longo dos segmentos de linha entre x_i e cada um dos vizinhos é escolhido aleatoriamente.

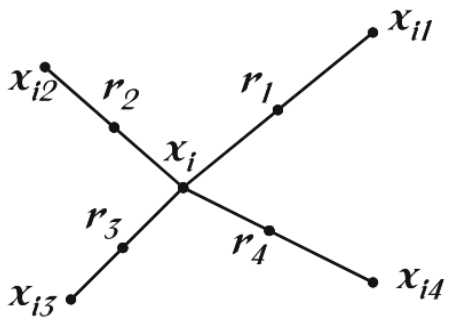


Figura 8 – Ilustração da geração de pontos sintéticos no algoritmo SMOTE. Fonte: Fernández et al. (2018).

O funcionamento formal do SMOTE pode ser resumido da seguinte forma (FERNÁNDEZ et al., 2018):

1. Define-se a quantidade total de sobreamostragem N necessária para aproximar uma distribuição balanceada (ex.: proporção 1:1).
2. Para cada instância p_i da classe minoritária:

a) Calculam-se seus k vizinhos mais próximos (VMPs) no espaço de atributos, considerando apenas exemplos da mesma classe.

b) Seleciona-se aleatoriamente um subconjunto desses k vizinhos.

c) Para cada vizinho selecionado, gera-se uma nova instância sintética s através da interpolação linear: $s = p_i + \lambda \cdot (p_{zi} - p_i)$, onde p_{zi} é o vizinho selecionado e λ é um número aleatório no intervalo $[0, 1]$.

A principal vantagem do SMOTE sobre a replicação simples é a capacidade de criar exemplos que generalizam o espaço de características da classe minoritária, em vez de apenas aumentar a ocorrência exata de pontos existentes. No entanto, como apontado por Fernández et al. (2018), o SMOTE pode apresentar algumas limitações, como a geração de exemplos sintéticos em regiões que causam sobreposição entre as classes e a possível criação de pontos ruidosos, especialmente se não houver consideração pela distribuição local dos dados.

Visando superar essas limitações, é usual combinar técnicas de *oversampling*, como o SMOTE, com métodos de *undersampling* ou limpeza de dados, formando abordagens híbridas. Um exemplo de combinação é o **SMOTE + Tomek Links**, em que aplica-se primeiro o SMOTE para gerar exemplos sintéticos e, em seguida, utiliza-se uma técnica de limpeza para remover instâncias ruidosas ou ambíguas de ambas as classes, visando obter agrupamentos de classes mais bem definidos (FERNÁNDEZ et al., 2018). Tais técnicas de pré-processamento funcionam em qualquer algoritmo de classificação, podendo ser facilmente incorporadas à construção de um modelo de aprendizado de máquina.

Análises de Componentes Principais

Para visualização do conjunto de dados de treinamento após a aplicação dos algoritmos de *oversampling* e *undersampling*, uma das ferramentas mais utilizadas é a Análise de Componentes Principais (do inglês *Principal Component Analysis* - PCA).

A ideia central da PCA, conforme Liu (2002), é reduzir a dimensionalidade de um conjunto de dados que possui um grande número de variáveis inter-relacionadas, mas mantendo o máximo possível da variação presente no conjunto original. Essa redução ocorre ao transformar as variáveis originais em um novo conjunto de variáveis, as Componentes Principais (CPs), que não são correlacionadas entre si. Essas componentes são ordenadas de forma que as primeiras retenham a maior parte da variação presente em todas as variáveis originais (LIU, 2002). Tecnicamente, o cálculo das componentes principais resume-se à solução de um problema de autovalores e autovetores de uma matriz simétrica, como a de covariância ou de correlação.

De acordo com Liu (2002), as primeiras q componentes principais definem o subespaço de dimensão q que melhor se ajusta aos dados, visando minimizar a soma das distâncias perpendiculares quadradas das observações originais até o subespaço. Essa medida, chamada de "bondade de ajuste" (*goodness-of-fit*), garante que a representação em duas dimensões seja a projeção que causa a menor distorção possível na configuração global dos dados.

Na Figura 9, apresenta-se o fundamento geométrico da PCA por meio da projeção ortogonal de um vetor bidimensional em um subespaço unidimensional. Assim, verifica-se como uma instância ou observação original (x_i) é projetada em um subespaço (representado pela reta z_1), resultando em um valor projetado (m_i) e em um vetor residual (r_i).

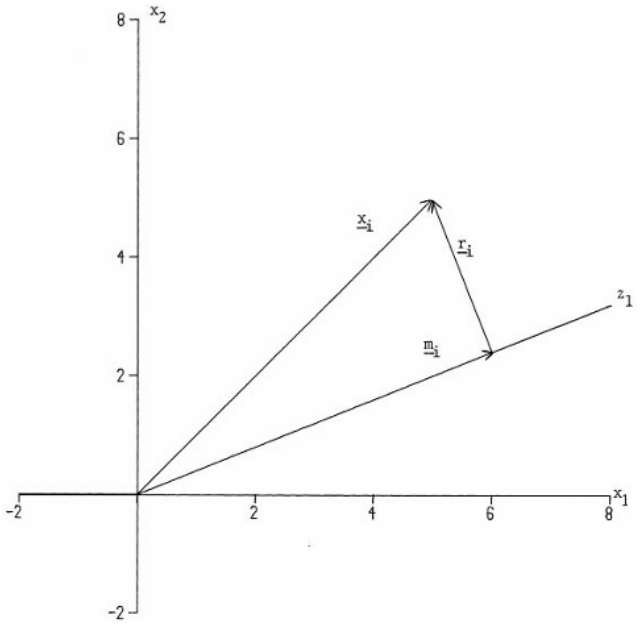


Figura 9 – Projeção ortogonal de um vetor bidimensional em um subespaço unidimensional. Fonte: Liu (2002).

Conforme Liu (2002), essa representação é a chave para entender a *goodness-of-fit* da técnica, uma vez que a PCA busca definir o subespaço de modo que a soma dos quadrados dessas distâncias perpendiculares (r_i) seja a menor possível.

2.3 Trabalhos Relacionados

Trabalhos recentes têm explorado o uso de técnicas de aprendizado de máquina para a classificação do teor de sílica no concentrado da flotação de minério de ferro. Nesse contexto, Ramos et al. (2025) propuseram um modelo de classificação baseado em *machine learning* interpretável para prever, em tempo real, se o teor de sílica no concentrado estaria acima ou abaixo do limite de $1,2\%$. O problema foi formulado como uma tarefa de classificação binária, distinguindo a classe crítica ($> 1,2\%$) da classe não crítica ($\leq 1,2\%$).

Ramos et al. (2025) desenvolveram o modelo a partir de dados industriais coletados ao longo de 18 meses de operação de uma planta de flotação de minério de ferro. Tais dados incluíram variáveis relacionadas às características da alimentação, às dosagens de reagentes e aos parâmetros operacionais. Para garantir interpretabilidade dos resultados, os autores adotaram o algoritmo *Explainable Boosting Machine* (EBM), permitindo a análise da influência individual das variáveis nas predições. Os resultados indicaram acurácia global de aproximadamente 76%,

com *recall* de 87% e *precision* de 73% para a classe crítica, o que evidencia o potencial da abordagem como ferramenta de suporte à tomada de decisão.

Os autores Jiang, Zhao e Liu (2023) estudaram a classificação qualitativa das condições de flotação a partir da análise de imagens da espuma. Nesse estudo, visando identificar diferentes estados operacionais do processo, propôs-se um método que combina imagens no espectro visível e no infravermelho próximo, por meio de técnicas de extração de características utilizadas em conjunto com algoritmos de aprendizado de máquina. Assim, as condições da espuma foram agrupadas em classes que representam os regimes de desempenho da flotação.

No estudo, Jiang, Zhao e Liu (2023) utilizaram dados experimentais coletados em uma planta de flotação. Com os dados, os autores testaram diferentes classificadores, incluindo máquinas de vetor de suporte (do inglês *Support Vector Machine* – SVM). Os resultados apresentaram melhor desempenho ao utilizar uma abordagem multimodal, ao invés de utilizar apenas um tipo de imagem. Assim, os autores alcançaram altos níveis de acurácia na classificação das condições da flotação.

Os autores Zhang e Gao (2021) desenvolveram um sensor virtual para a classificação qualitativa do grau da espuma de flotação de minério de ferro a partir de imagens do processo. Os autores formularam o problema como uma tarefa de classificação multiclasse, na qual o grau da espuma foi dividido em sete categorias qualitativas, variando de muito baixo a muito alto.

Nesse caso, o método proposto baseou-se em uma arquitetura híbrida de aprendizado profundo, que combina três redes neurais convolucionais profundas, sendo elas AlexNet, ResNet101 e ShuffleNet. Dessa forma, os autores Zhang e Gao (2021) utilizaram cada sub-rede para extrair automaticamente as características das imagens de espuma. Após, as características foram utilizadas para a classificação multiclasse, sendo que a avaliação do modelo em cenário industrial alcançou acurácia de 97% na identificação das classes qualitativas.

De forma semelhante, Silva e Galéry (2013) investigaram o uso da análise de imagens da espuma como indicador qualitativo do teor de sílica no concentrado de uma flotação reversa de minério de ferro, baseando-se no tamanho das bolhas da espuma. Nesse estudo, os autores empregaram o espectro de textura, construído a partir do histograma de frequência das unidades de textura identificadas na imagem. O valor do pico central do histograma (Mid_TU) foi utilizado como parâmetro representativo do tamanho médio das bolhas, o que permitiu o agrupamento das imagens em classes qualitativas, definidas como bolhas pequenas, médias e grandes.

Silva e Galéry (2013) conduziram o estudo com o uso de dados industriais provenientes de colunas de flotação da Samarco Mineração S.A.. Após a definição das classes de tamanho de bolhas, os autores avaliaram a relação entre essas classes e o teor de sílica no concentrado por meio de testes estatísticos, sendo eles ANOVA e Kruskal–Wallis. Os resultados indicaram diferenças estatisticamente significativas entre as classes, sendo observado que imagens associadas a bolhas maiores apresentaram, em média, menores teores de sílica e menor variabilidade no concentrado. Dessa forma, os autores evidenciaram o potencial da classificação qualitativa da espuma como suporte à avaliação do desempenho do processo de flotação.

Por outro lado, há também estudos recentes voltados para modelos de regressão na flotação. Zhang, Gao e Qi (2022) utilizaram as características de cor da imagem da espuma para desenvolver um sensor virtual capaz de prever o teor de rejeitos na espuma da flotação de minério de ferro. Ainda, ao utilizarem a tecnologia *digital twin* e imagens da espuma, Zhang e Gao (2022) projetaram um sensor virtual para prever o grau de rejeitos no concentrado da flotação de minério de ferro, com o objetivo de alimentar a entrada de um controlador de reagentes, aumentando a eficiência da produção e evitando o desperdício de reagentes.

Ainda, Pural (2023) utilizou dados como porcentagem de ferro e sílica na alimentação, vazão de reagentes, nível das colunas, teor de sílica no concentrado, entre outras variáveis de uma flotação de minério de ferro, para desenvolver um sensor virtual de previsão do teor de silicatos no concentrado. Em Liu et al. (2020), variáveis de entrada como grau de alimentação, de concentrado, de rejeitos e dosagens de reagentes foram selecionadas para prever tanto o teor do concentrado quanto o teor de rejeitos em uma flotação de minério de ferro.

O autor Guedes (2020) desenvolveu um sensor virtual para estimar, em tempo real, o teor de ferro no concentrado de uma flotação de minério de ferro. Para isso, foram utilizadas variáveis de processo coletadas ao longo da operação industrial. Na etapa de modelagem, foram empregadas Redes Neurais Artificiais (RNA) e o método *Random Forest*, o que permitiu a comparação de desempenho entre diferentes abordagens de aprendizado de máquina para a predição do teor.

3 Materiais e Métodos

Neste capítulo, apresentam-se os materiais utilizados e a metodologia proposta para o desenvolvimento do sensor virtual.

3.1 Dados

A base de dados utilizada neste trabalho possui os dados do processo de flotação da usina de beneficiamento da Mina de Timbopeba, da Vale S.A., localizada em Ouro Preto, estado de Minas Gerais, Brasil.

Os registros históricos da base foram obtidos a partir do Sistema de Gestão de Informações sobre Processos (PIMS), por meio do suplemento PI DataLink para Microsoft Excel, que foi utilizado como interface de extração dos dados armazenados no produto PI System^{TM}, da fabricante AVEVA (AVEVA). Tais registros incluem dados históricos de variáveis medidas por sensores e de variáveis cujos valores são provenientes de análises laboratoriais.

3.2 Metodologia

Nesta seção são descritos os procedimentos adotados para o desenvolvimento do sensor virtual proposto. Todas as etapas metodológicas foram realizadas em ambiente Python com o auxílio de bibliotecas como `pandas`, `numpy`, `scipy`, `matplotlib`, `sklearn` e `seaborn`.

Nesse sentido, realizaram-se o tratamento da base de dados, a análise estatística das variáveis e o cálculo de correlações. Além disso, realizou-se a preparação da base final utilizada no treinamento e no teste dos modelos de classificação.

3.2.1 Pré-processamento dos dados

Na usina de beneficiamento da Mina de Timbopeba, identificaram-se 109 *tags* relacionadas ao processo de flotação de minério de ferro, tais como variáveis de medição, estado de funcionamento dos equipamentos, parâmetros de processo, dados de análises laboratoriais e informações provenientes de imagens (velocidade e RGB das bolhas).

Após a identificação das *tags*, levantaram-se os dados referentes à cada uma delas, no período de 22/11/2024 a 26/07/2025, com uma frequência de amostragem de 20 segundos. A aquisição e o alinhamento dos dados, de modo que todos estivessem representados em um mesmo intervalo de amostragem, foram efetuados pelo sistema PIMS disponível na unidade operacional.

Em seguida, realizou-se uma inspeção inicial dos dados, assim como entrevistas com o operador do processo, a fim de identificar *tags* congeladas, com medições inconsistentes, dados faltantes ou cujas informações não fossem relevantes para o resultado dos concentrados da

flotação. Sendo assim, das 109 *tags* iniciais, consideraram-se 43, levando-se em consideração as avaliações mencionadas.

Seleção de variáveis

Inicialmente, com base em estudos relacionados e em entrevistas com o operador da usina, selecionaram-se as variáveis de interesse para o estudo, com destaque para o teor de sílica no concentrado, definido como variável alvo. As variáveis de entrada incluíram medições de vazão de alimentação da polpa, dosagens de amido e amina, níveis das células de flotação, pH, etc., além de dados de resultados laboratoriais, como granulometrias, teor de ferro e sílica na alimentação e no rejeito, entre outros.

Além disso, desconsideraram-se as variáveis medidas pelas câmeras, visto que estas apresentaram inconsistências de medição de velocidades e insuficiência de dados RGB, sendo necessário realizar ajuste do sistema de medição e de coleta das câmeras antes que os dados possam ser utilizados.

Tratamento de dados faltantes

Na base de dados, realizou-se um tratamento a fim de limpar dados com valores inválidos. Sendo assim, excluíram-se todas as linhas em que pelo menos uma variável contivesse valores "Scan Off" ou "Bad", que são gerados quando há, por exemplo, falha de comunicação com o CLP.

Remoção de dados por condição de processo

Verificou-se, com intermédio do operador da planta e por meio de análise dos dados que, no processo de flotação estudado, considera-se parada da planta quando a vazão de alimentação está abaixo de 500 m^3/h. Na Figura 10, apresenta-se a porcentagem de tempo em que os motores das células de flotação não estiveram em funcionamento durante o período de 29/11/2024 a 04/12/2024, maior período identificado em que a vazão de alimentação esteve abaixo de 500 m^3/h, comprovando parada da planta nesse caso.

Por esse motivo, para capturar apenas momentos ativos do processo, excluíram-se as linhas em que a variável de vazão de alimentação estava abaixo desse valor.

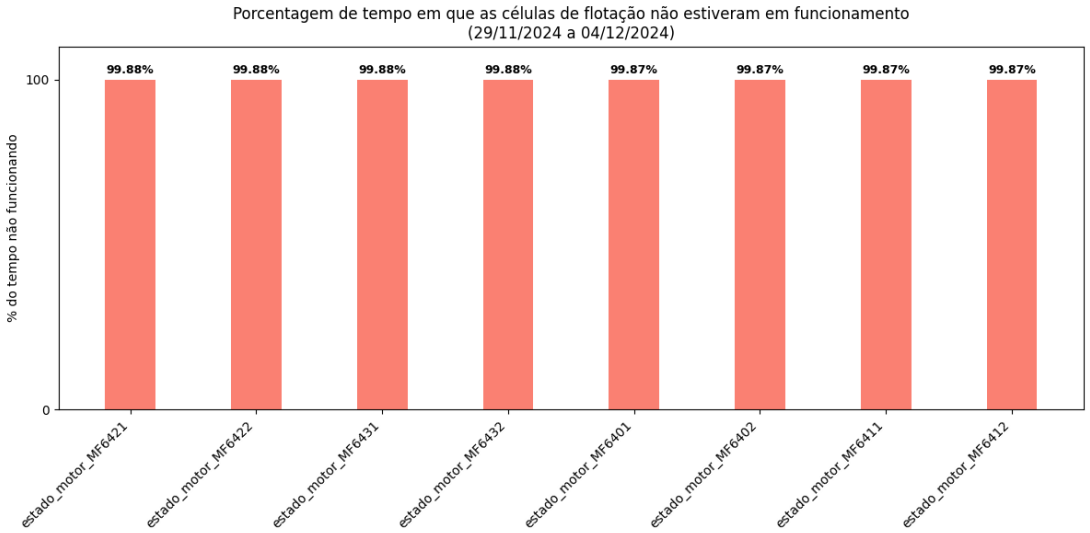


Figura 10 – Estado dos motores das células quando a vazão está abaixo de 500 m/h.

Consolidação de variáveis relacionadas

Na planta de flotação analisada, são realizadas medições de pH em ambas as células *rougher*. Apresenta-se, na Figura 11, o diagrama P&ID do processo, em que os valores obtidos (AIT) passam por um procedimento de comparação (AY), no qual é selecionado o maior entre os dois. Isso ocorre para evitar redundância na leitura, uma vez que as primeiras células *rougher* recebem a mesma alimentação, além de ser importante para capturar o pior caso, pois podem haver problemas na leitura do sensor. Esse valor é então utilizado como referência para o controle regulatório de pH da operação. Dessa forma, aplicou-se o mesmo tratamento à base de dados, consolidando os valores apenas em uma coluna que representa o maior pH entre as duas medições.

Outra variável consolidada foi a sílica na alimentação. Isso porque o circuito de flotação recebe alimentação tanto da mina de Timbopeba quanto da Mina de Fábrica Nova, também pertencente à Vale S.A.. Dessa forma, considerou-se, para análise, uma média simples entre as duas variáveis, por sugestão do operador da planta.

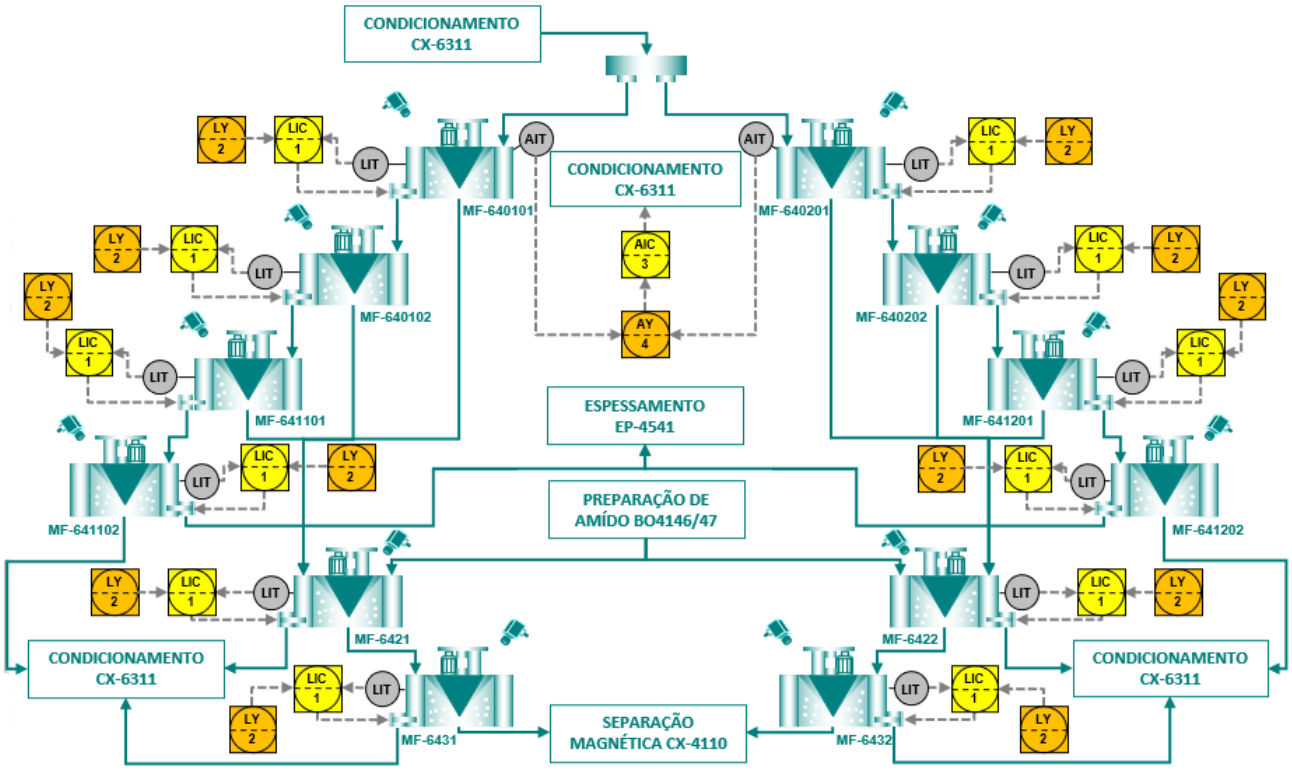


Figura 11 – Diagrama P&ID do controle do processo de flotação.

Tratamento para variáveis de laboratório

Na planta de flotação estudada, existem variáveis que são provenientes de análises laboratoriais, como os teores de sílica e de ferro na alimentação, no concentrado e no rejeito, além das granulometrias dos materiais. Diferentemente das variáveis de processo, que são medidas continuamente por sensores, os dados laboratoriais são atualizados apenas a cada 2h, quando um novo resultado é disponibilizado no sistema PIMS. Durante o intervalo entre coletas, os valores permanecem constantes na base de dados, que possui amostragem de 20 em 20 segundos.

Na Tabela 1, apresenta-se a relação entre o período avaliado e o resultado da análise laboratorial. Nesse sentido, quando um novo valor é registrado no PIMS — por exemplo, às 02:00h — ele representa o comportamento médio do processo entre 00:00h e 01h40, sendo que os 20 minutos antes do lançamento representam o tempo de transporte da amostra até o laboratório. Ressalta-se que a análise da amostra demanda tempo, sendo assim, nesse exemplo, a análise fica pronta algum tempo depois de 02:00h. No entanto, os profissionais responsáveis pelo laboratório lançam o resultado da amostra no PIMS como se não houvesse atraso.

Para que as variáveis laboratoriais pudessem ser utilizadas como entradas contínuas no modelo do sensor virtual, realizou-se o cálculo de uma média móvel deslocada. Esse procedimento teve como objetivo fornecer um valor representativo de forma constante, compensando o fato de que os dados laboratoriais são disponibilizados com atraso. Na operação, as amostras são coletadas automaticamente a cada 10 minutos por amostradores do processo, mas passam por etapas de análise antes de serem lançadas no sistema PIMS.

Por esse motivo, adotou-se uma estratégia em que, a cada nova mudança de valor da

Tabela 1 – Relação entre período avaliado e resultado da análise laboratorial

Dados PIMS	**Resultado Laboratorial**				
	02:00	**04:00**	**06:00**	**08:00**	**10:00**
00:00 – 01:40					
02:00 – 03:40					
04:00 – 05:40					
06:00 – 07:40					
08:00 – 09:40					

variável de análise laboratorial, calcula-se a média dos três últimos valores, atribuindo esse resultado como valor de entrada até que uma nova atualização seja registrada. A definição da média móvel com os três últimos valores decorreu de recomendação do operador da planta, que observou que esse intervalo é capaz de representar, de forma adequada, a tendência dos resultados laboratoriais, preservando a sensibilidade às variações do processo. As médias móveis de cada variável foram, então, armazenadas em colunas auxiliares, permitindo que o modelo tenha acesso a uma estimativa contínua e mais realista do comportamento do processo.

Consolidação dos dados por amostragem do produto

Para consolidar os dados de processo em alinhamento com a frequência das variáveis laboratoriais, adotou-se uma estratégia baseada na mediana dos valores registrados em janelas temporais específicas. A escolha da mediana, em detrimento da média, deve-se à sua maior robustez frente a variações pontuais e à sua capacidade de representar melhor o comportamento predominante das variáveis ao longo do tempo, conforme evidenciado na Figura 12, em que apresenta-se a variação da densidade de alimentação do processo de flotação no período de 24/11/2024 a 27/11/2024.

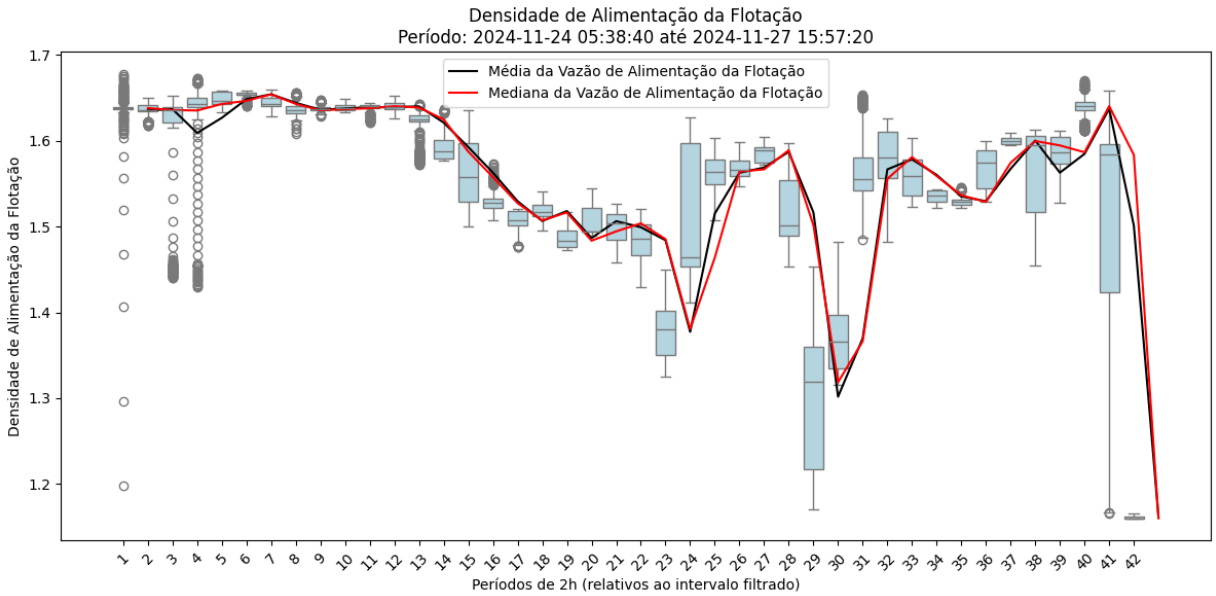


Figura 12 – Densidade de alimentação da flotação.

A base de dados final foi construída a partir da consolidação das variáveis por meio de medianas, em janelas temporais de 2h, em razão desse tempo ser equivalente ao intervalo entre lançamentos, no PIMS, dos valores percentuais de sílica no concentrado. Sendo assim, desconsideraram-se intervalos de tempo cujo valor percentual de sílica no concentrado permanecia estático por mais de 2h, uma vez que tais intervalos podem indicar parada de processo ou falha no lançamento de valores no PIMS. Esse procedimento garantiu que os dados utilizados para treinamento representassem condições estáveis do processo. O resultado foi uma base de dados consolidada, contendo variáveis de entrada tratadas e a variável alvo devidamente ajustada, pronta para ser utilizada nos modelos de classificação.

3.2.2 Análise estatística

Após o tratamento da base de dados, obtiveram-se as possíveis variáveis de processo a serem utilizadas como entradas para os modelos de classificação. A Tabela 2 apresenta a seleção de variáveis para a realização de testes dos modelos.

Com a base tratada e as variáveis selecionadas, aplicaram-se análises estatísticas para investigar a relação entre as variáveis de processo e a variável alvo.

Tabela 2 – Variáveis de processo utilizadas na análise

Variável	**Unidade**
Nível de polpa na célula de flotação 6401-01	mm
Nível de polpa na célula de flotação 6401-02	mm
Nível de polpa na célula de flotação 6402-01	mm
Nível de polpa na célula de flotação 6402-02	mm
Nível de polpa na célula de flotação 6411-01	mm
Nível de polpa na célula de flotação 6411-02	mm
Nível de polpa na célula de flotação 6412-01	mm
Nível de polpa na célula de flotação 6412-02	mm
Nível de polpa na célula de flotação 6421-01	mm
Nível de polpa na célula de flotação 6422-01	mm
Nível de polpa na célula de flotação 6431-01	mm
Nível de polpa na célula de flotação 6432-01	mm
Parâmetro de dosagem de amido no processo de flotação	g/t
pH na flotação	–
Densidade da polpa na alimentação da flotação	t/m^3
Vazão volumétrica da alimentação da flotação	m^3/h
Vazão de amina na flotação	L/h
Dosagem específica de amina na flotação	g/t
Vazão mássica de sólidos na alimentação	t/h
Vazão mássica de amido na flotação	kg/h

3.2.3 Seleção, treinamento e validação do modelo

Após o pré-processamento e a análise estatística das variáveis, definiram-se os procedimentos para seleção, treinamento e validação dos modelos de classificação.

Nos modelos de classificação, utilizaram-se como entradas o pH, as vazões de alimentação, de amina, de amido, de sólidos na alimentação, bem como a dosagem de amina. Além disso, utilizaram-se as variáveis de nível, pois estas aumentaram o desempenho dos modelos. No caso das variáveis calculadas com as médias móveis, utilizaram-se a granulometria passante da peneira (-0,045 mm), a granulometria acumulada na peneira (+0,15 mm), sílica e ferro no rejeito, sílica e ferro na alimentação e a variável de sílica média na alimentação.

Estruturou-se a estratégia de modelagem de forma sequencial. Primeiramente, utilizou-se a classificação binária para avaliar o comportamento do *Random Forest* diante de uma situação de menor complexidade. Após a análise de desempenho deste primeiro estágio, expandiu-se o modelo para uma estrutura multiclasse, com o objetivo de aumentar a quantidade de cenários capazes de serem identificados e fornecer uma resposta mais refinada a ser implementada nos sistemas de controle avançado. Tal abordagem permite que o sensor virtual atue em mais faixas de teor de sílica.

Assim, criaram-se duas bases de dados distintas para a realização de testes de ambos os modelos:

- **Classificação binária**: categorizou-se o teor de sílica no concentrado em duas classes, "Normal" (valores especificados em até 2,5%, a partir de entrevistas com a operação) e "Alta" (valores acima de 2,5%).
- **Classificação multiclasse**: dividiram-se os rótulos em três faixas, sendo elas "baixo" (valores abaixo de 1,5%), "moderado" (valores entre 1,5% e 2,5%) e "alto" (valores acima de 2,5%), o que permitiu maior granularidade na análise dos resultados.

Nas análises dos dados a serem utilizados para as classificações binária e multiclasse, observou-se desbalanceamento entre as categorias. Nas Figuras 13 e 14 apresentam-se, respectivamente, o balanceamento das bases finais para ambas as classificações, sendo que cada uma delas possui uma quantidade total de 1909 amostras.

Inicialmente, buscando mitigar o impacto do desbalanceamento entre as classes na etapa de treinamento do modelo, aplicaram-se estratégias de balanceamento, incluindo o método *balanced* da biblioteca `sklearn` e de pesos calculados conforme o balanceamento original da base de dados.

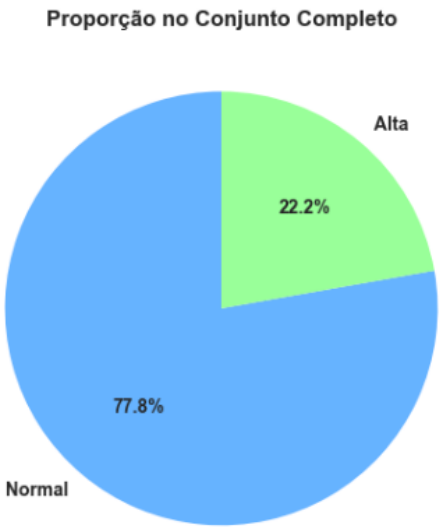


Figura 13 – Proporção na base de dados final para classificação binária.

Figura 14 – Proporção na base de dados final para classificação multiclasse.

Na estratégia de pesos calculados, dada pela Equação 3.1, verificou-se a quantidade de amostras de cada classe na base de dados e dividiu-se cada uma pela quantidade de amostras da classe predominante. Dessa forma, a classe predominante recebe peso unitário, enquanto as classes minoritárias recebem pesos maiores, proporcionais ao inverso de sua frequência relativa na base de dados, mitigando o impacto do desbalanceamento durante o treinamento do modelo.

$$w_i = \frac{N_{\max}}{N_i}, \tag{3.1}$$

em que w_i é o peso associado à classe i, N_i é o número de amostras da classe i e $N_{\max}$ é o número de amostras da classe predominante, definido como $N_{\max} = \max(N_1, N_2, \ldots, N_k)$.

Para balancear os dados de forma mais robusta, utilizaram-se, em conjunto, as técnicas SMOTE e Tomek Links. Tais técnicas formam um método híbrido que combina as estratégias de *undersampling* e *oversampling*.

Em seguida, desenvolveram-se os modelos de classificação utilizando o algoritmo *Random Forest Classifier*, com divisão dos dados em treino (70%) e teste (30%). Realizou-se a validação do modelo por meio de métricas como acurácia, precisão, *recall* e *F1-score*, além da análise de matrizes de confusão.

4 Resultados

Neste capítulo, apresentam-se os resultados obtidos com os modelos de classificação desenvolvidos para a previsão do teor de sílica no concentrado da flotação. Os experimentos foram realizados com dados históricos tratados e consolidados, conforme descrito no Capítulo 3.

Programaram-se os modelos em Python, versão 3.9.13, por meio do ambiente de desenvolvimento Visual Studio Code, editor de código *open source*, em um computador com processador Intel® Core™ i7, 8 GB de memória RAM, arquitetura x64 e sistema operacional Windows 11.

4.1 Modelo de Classificação

Nesta seção, apresentam-se os resultados para o modelo de classificação utilizando o algoritmo *Random Forest Classifier* para a base de dados tratada da usina Timbopeba, da Vale S.A., Brasil.

4.1.1 Classificação em duas faixas

Primeiramente, objetivou-se categorizar o teor de sílica em duas faixas de qualidade. Na Figura 15, apresenta-se a distribuição das classes após a divisão de dados em treino e teste.

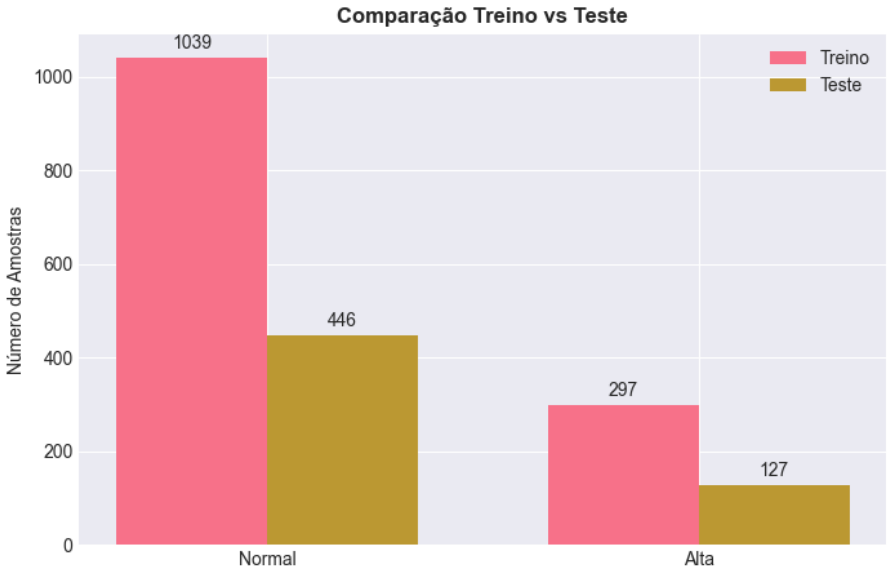


Figura 15 – Proporção de dados entre as classes nas bases de treinamento e teste do modelo de duas faixas de classificação

Em razão do desbalanceamento, utilizou-se a técnica de balanceamento por pesos na etapa de treinamento do modelo. Observa-se, na Figura 16, o resultado na matriz de confusão. Nesse caso, os acertos estão destacados na cor verde e os erros na cor vermelha.

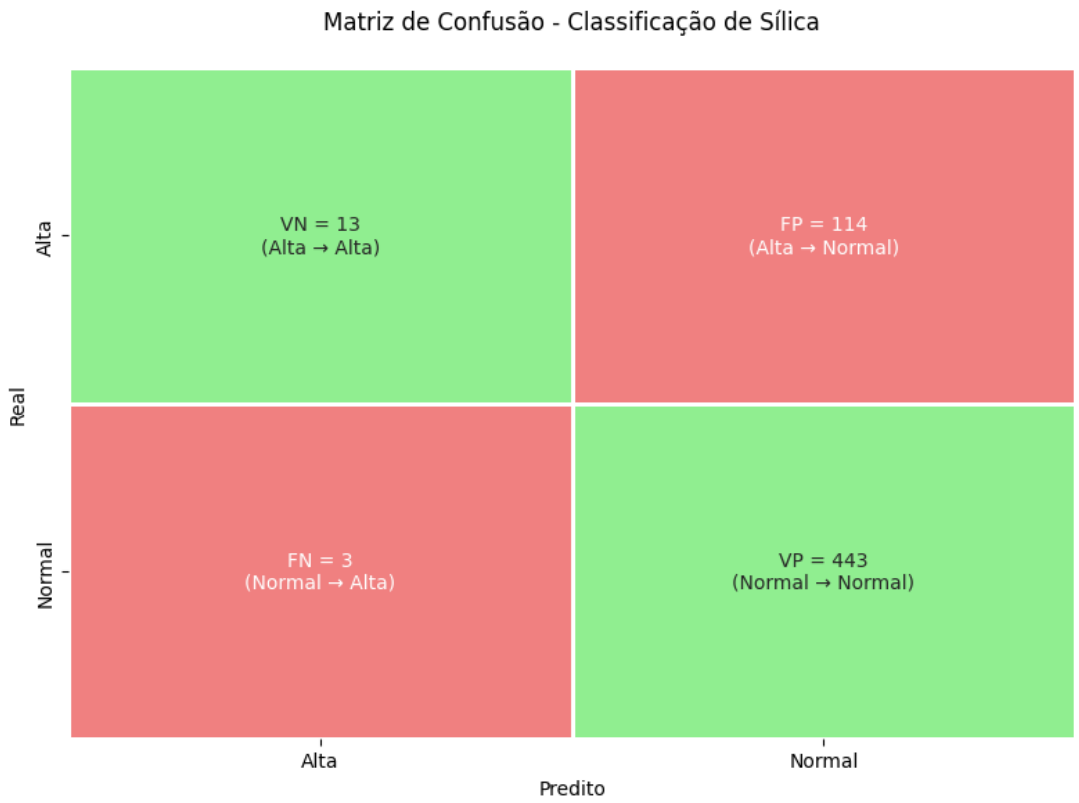


Figura 16 – Matriz de confusão para duas faixas de classificação

A Tabela 3 apresenta o relatório de classificação para este modelo.

Tabela 3 – Relatório de classificação para 2 classes

Classe	**Precisão**	**Recall**	**F1-score**	**Suporte**
Alta	0.81	0.10	0.18	127
Normal	0.80	0.99	0.88	446
Acurácia			0.80	
Média Macro	0.80	0.55	0.53	573
Média Ponderada	0.80	0.80	0.73	573

Tanto a Tabela 3 quanto a matriz de confusão obtidas para a classificação da sílica em duas faixas evidenciam um comportamento assimétrico do classificador entre as classes analisadas.

Para a classe "Normal", observa-se um desempenho elevado, com *recall* de $0,99$ e *precision* de 0,80. Isso indica que praticamente todas as amostras realmente pertencentes à classe Normal foram corretamente identificadas pelo modelo, conforme evidenciado pelos 443 verdadeiros positivos e apenas 3 falsos negativos na matriz de confusão. O elevado valor de *f1-score* (0,88) reforça a consistência do modelo para essa classe.

Por outro lado, a classe "Alta" apresenta desempenho significativamente inferior, especialmente em termos de *recall*, que foi de apenas 0,10. Esse resultado indica que o modelo tem grande dificuldade em identificar corretamente amostras de sílica nesse faixa, classificando incorretamente a maioria delas como "Normal". Tal comportamento é corroborado pela matriz

de confusão, na qual se observam 114 falsos positivos (amostras reais da classe Alta classificadas como Normal) e apenas 13 verdadeiros positivos. Apesar disso, a *precision* da classe Alta foi de 0,81, o que indica que, quando o modelo prevê a classe Alta, essa previsão tende a estar correta, embora ocorra com baixa frequência.

Nesse modelo, tem-se uma acurácia global de 0,80, valor influenciado pelo bom desempenho na classe "Normal", que representa a maioria das amostras. Esse desbalanceamento entre as classes impacta diretamente as métricas agregadas, como o *macro average*, cujo *recall* foi de 0,55, refletindo a baixa sensibilidade do modelo para a classe minoritária.

Portanto, embora o modelo apresente desempenho satisfatório para a identificação da condição "Normal", não há uma detecção confiável de níveis elevados de sílica, o que é crítico no processo de flotação.

Com o objetivo de aprimorar as métricas de classificação para duas faixas, adotou-se um modelo de *Random Forest Classifier* associado a uma estratégia de balanceamento híbrido. Essa abordagem combina as técnicas SMOTE e Tomek Links, disponibilizadas pela biblioteca `imbalanced-learn`. O balanceamento é aplicado exclusivamente na etapa de treinamento, o que promove uma distribuição equitativa entre as classes apenas durante o processo de aprendizagem do modelo. Na Figura 17, apresenta-se o balanceamento resultante da aplicação destas técnicas conjuntas. Dessa maneira, assegura-se um treinamento balanceado, que preserva a distribuição original dos dados no conjunto de teste, permitindo uma avaliação mais realista do desempenho do classificador.

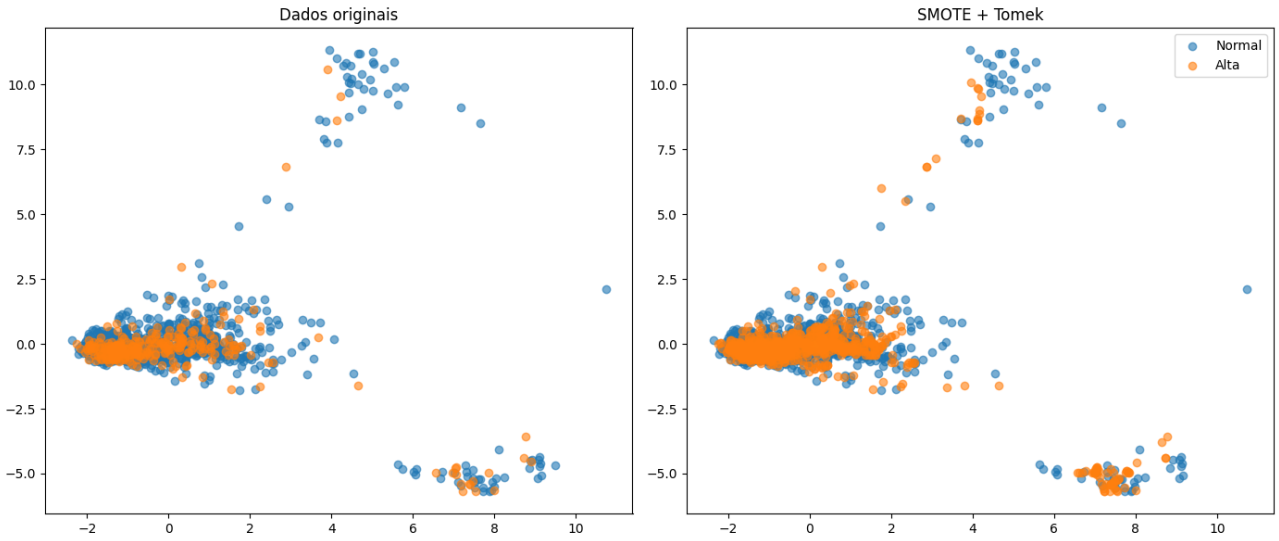


Figura 17 – Balanceamento de dados no treinamento após a aplicação de SMOTE + Tomek Links

Na Figura 18, apresenta-se a matriz de confusão resultante dessa estratégia. Na Tabela 4, apresenta-se o relatório de classificação para esse caso.

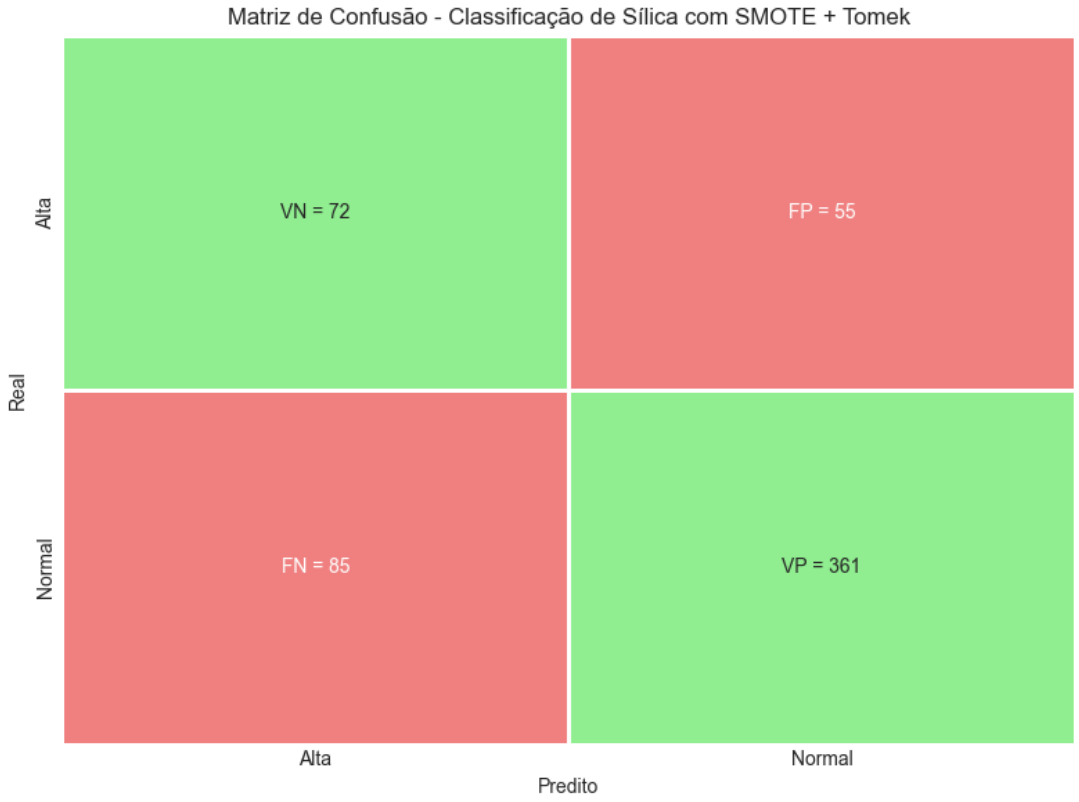


Figura 18 – Matriz de confusão para duas faixas de classificação utilizando técnica híbrida de balanceamento

Tabela 4 – Relatório de classificação do modelo *Random Forest* com balanceamento híbrido (SMOTE + Tomek)

Classe	**Precisão**	**Recall**	**F1-score**	**Suporte**
Alta	0.46	0.57	0.51	127
Normal	0.87	0.81	0.84	446
Acurácia		0.76		
Média Macro	0.66	0.69	0.67	573
Média Ponderada	0.78	0.76	0.76	573

Tais resultados revelam uma melhoria substancial na capacidade discriminativa do classificador. Ainda que a acurácia global tenha apresentado uma leve redução para $0,76$, observa-se um aumento expressivo no *recall* da classe *Alta*, que passou de $0,10$ para $0,57$, refletindo uma identificação mais equilibrada entre as classes. Adicionalmente, as métricas de média macro (precisão, *recall* e F1-score) apresentaram ganhos consistentes, indicando uma distribuição mais uniforme do desempenho do modelo entre as classes.

Além disso, observa-se uma redução no *recall* da classe "Normal", que diminuiu de $0,99$ para $0,81$, e no *F1-score*, que passou de $0,88$ para $0,84$. Isso porque, tornando o modelo mais sensível para a identificação da classe "Alta", ocorreu um aumento na quantidade de amostras da classe "Normal" classificadas incorretamente como pertencentes à classe minoritária, elevando o número de FPs da classe "Alta". Tal efeito é esperado em cenários de balanceamento, uma vez que o modelo passa a reduzir o viés em favor da classe majoritária e assume um

comportamento mais conservador. Ainda assim, considerando que a classe “Alta” representa a condição operacional mais crítica no processo de flotação, o ganho em sua capacidade de detecção justifica a leve perda de desempenho na classe ”Normal”.

4.1.2 Classificação em três faixas

Esta seção apresenta os resultados para a classificação com 3 faixas. Na Figura 19, apresenta-se a distribuição dos dados separados para as etapas de treino e teste do modelo.

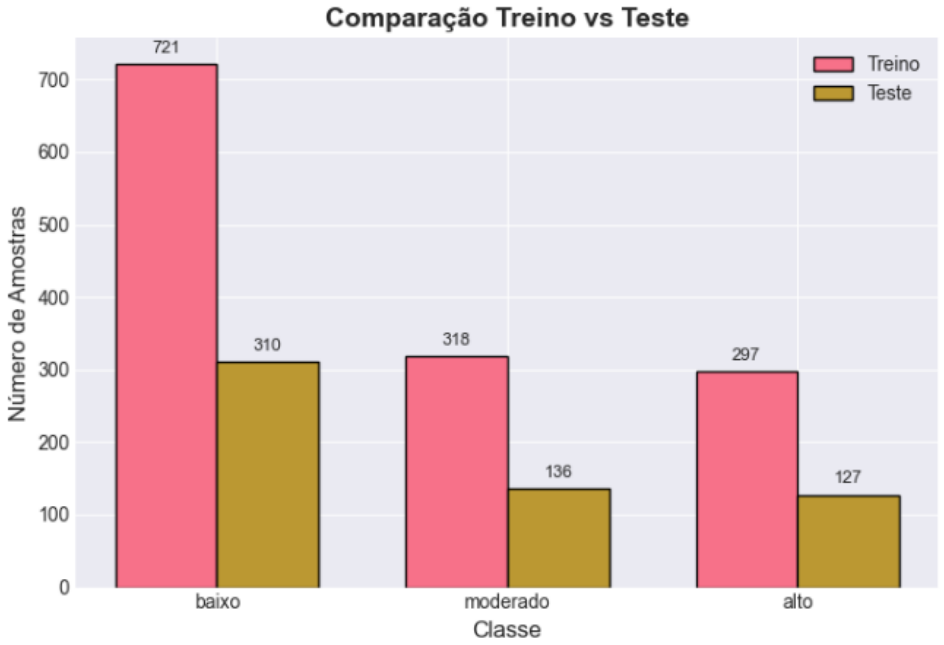


Figura 19 – Proporção de dados entre as classes nas bases de treinamento e teste do modelo de três faixas de classificação

Com o treinamento do modelo utilizando a estratégia de balanceamento por pesos, obteve-se a matriz de confusão apresentada na Figura 20, onde os acertos estão indicados em verde e os erros em vermelho.

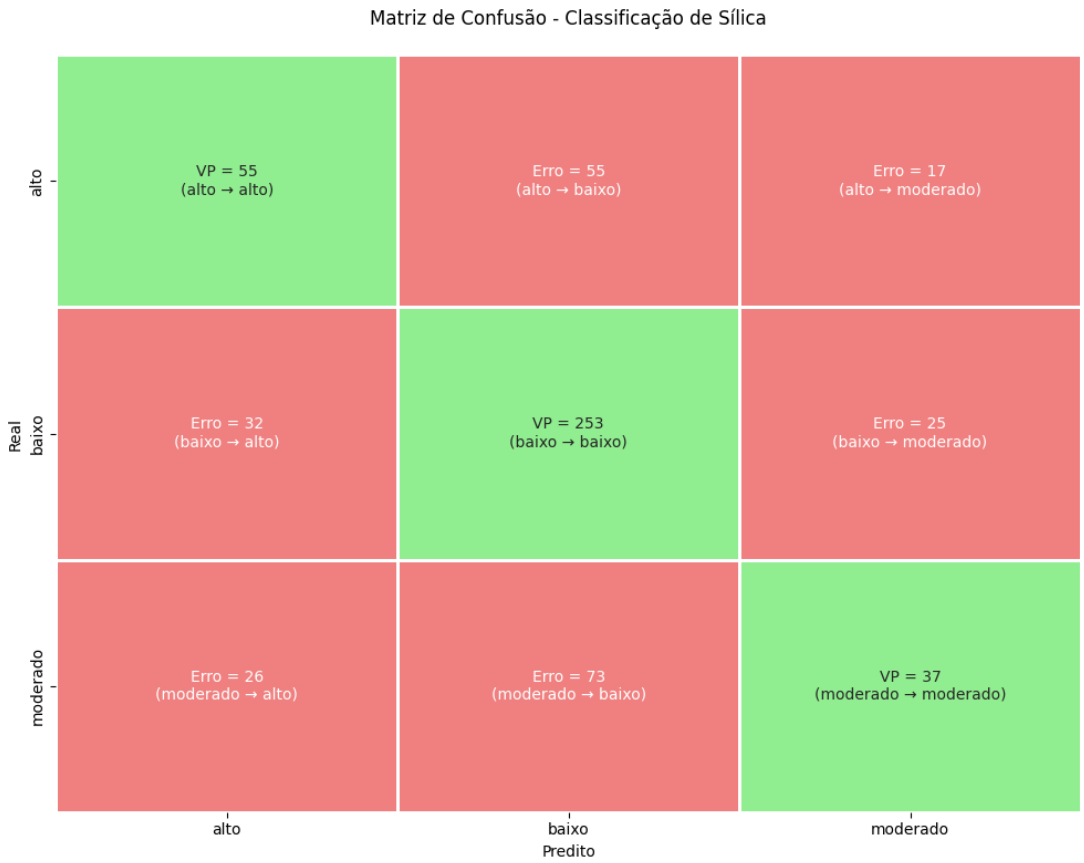


Figura 20 – Matriz de confusão para três faixas de classificação

Na Tabela 5, apresenta-se o relatório de classificação do modelo utilizando-se 3 faixas para o teor de sílica no concentrado da flotação (baixo, moderado e alto). Nesse caso, observa-se uma acurácia de 60%, indicando desempenho moderado para a classificação multiclasse.

Tabela 5 – Relatório de Classificação para Três Classes

Classe	**Precisão**	**Recall**	**F1-score**	**Suporte**
alto	0.49	0.43	0.46	127
baixo	0.66	0.82	0.73	310
moderado	0.47	0.27	0.34	136
Acurácia		0.60		
Média Macro	0.54	0.51	0.51	573
Média Ponderada	0.58	0.60	0.58	573

A avaliação desse modelo revela um desempenho global inferior quando comparado ao cenário com duas classes, evidenciando o aumento da complexidade do problema. A classe "baixo" apresentou o melhor desempenho entre as três, com *recall* de 0,82 e *precision* de 0,66. Esse resultado indica que a maior parte das amostras realmente pertencentes a essa faixa foram corretamente identificadas pelo modelo. A matriz de confusão confirma esse comportamento, com 253 amostras corretamente classificadas como baixo, embora ainda haja confusões relevantes principalmente com as classes alto e moderado. O *f1-score* de 0,73 reforça a maior robustez do classificador para essa classe.

Para a classe “alto”, observa-se um desempenho intermediário, com *precision* de 0,49 e *recall* de 0,43. Isso indica que menos da metade das amostras reais dessa classe foram corretamente identificadas. A matriz de confusão evidencia uma dispersão significativa das amostras da classe “alto”, que são frequentemente confundidas com as classes “baixo” e “moderado”, o que reduz a sensibilidade do modelo para a detecção de níveis elevados de sílica.

A classe “moderado” apresentou o pior desempenho geral, com *recall* de apenas 0,27 e *f1-score* de 0,34. Esse resultado indica grande dificuldade do modelo em identificar corretamente essa faixa, classificando a maioria das amostras moderadas como “baixo” ou “alto”.

As métricas de *macro average* (recall de 0,51) evidenciam que o desempenho médio entre as classes é limitado, especialmente devido à baixa sensibilidade para a classe “moderado”. Já o *weighted average* reflete a influência da classe “baixo”, que possui maior número de amostras e melhor desempenho relativo.

De forma geral, os resultados indicam que, embora o modelo consiga identificar razoavelmente bem a faixa baixa do teor de sílica no concentrado, ele apresenta dificuldades significativas na separação entre as faixas “alto” e “moderado”.

Com o objetivo de melhorar a detecção das classes minoritárias, aplicou-se, também, nesta base de dados, a estratégia híbrida de balanceamento. O resultado do balanceamento está apresentado na Figura 21. Além disso, apresentam-se os resultados do modelo na Figura 22 e na Tabela 6.

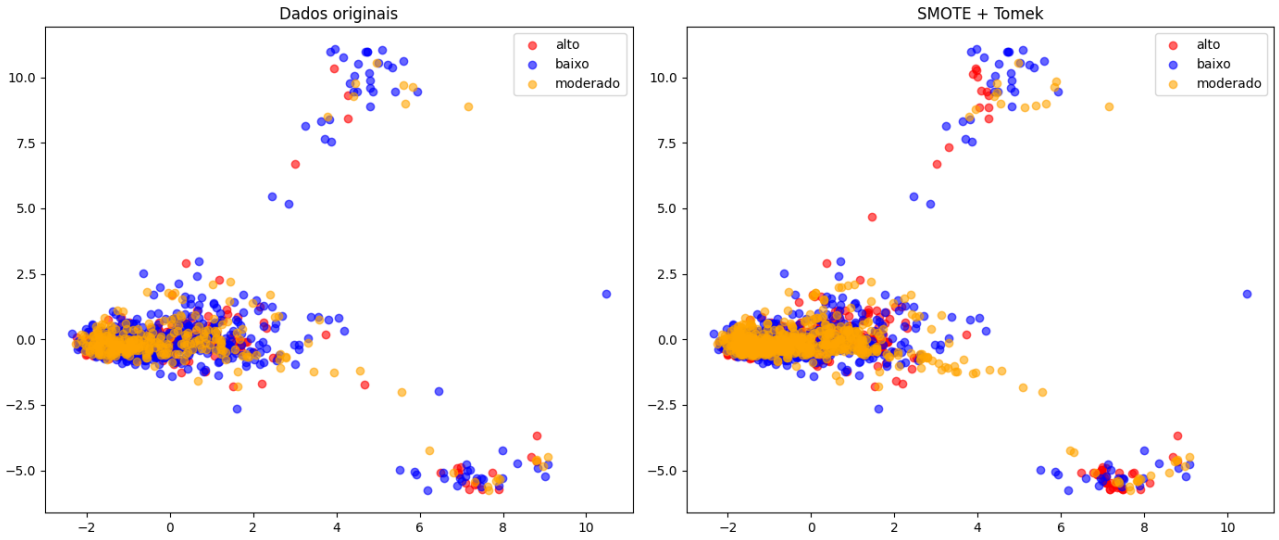


Figura 21 – Balanceamento de dados no treinamento após a aplicação de SMOTE + Tomek Links

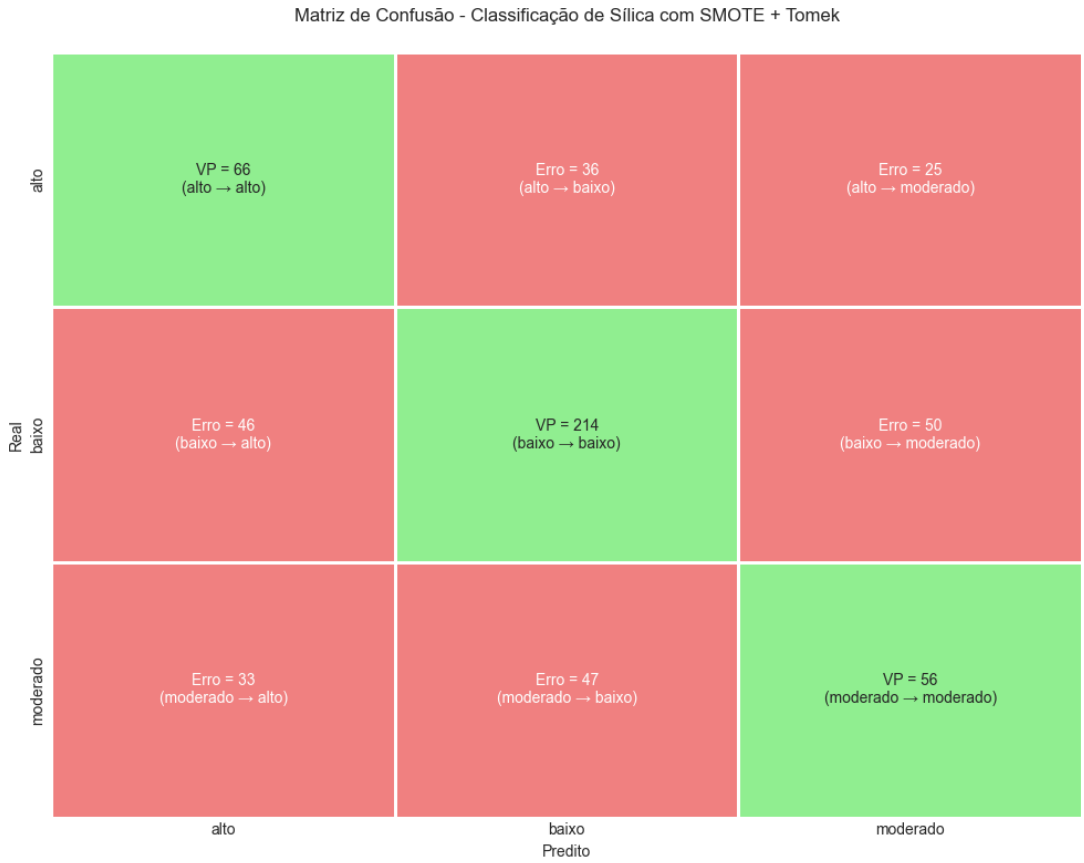


Figura 22 – Matriz de confusão para três faixas de classificação utilizando técnica híbrida de balanceamento

Tabela 6 – Relatório de classificação do modelo *Random Forest* com balanceamento híbrido (SMOTE + Tomek) para três classes

Classe	**Precisão**	**Recall**	**F1-score**	**Suporte**
alto	0.46	0.52	0.49	127
baixo	0.72	0.69	0.71	310
moderado	0.43	0.41	0.42	136
Acurácia			0.59	
Média Macro	0.53	0.54	0.54	573
Média Ponderada	0.59	0.59	0.59	573

Os resultados evidenciam uma distribuição mais equilibrada das métricas entre as classes. Observa-se um aumento no *recall* da classe *moderado*, que passa de $0,27$ para $0,41$, bem como uma melhora na classe "alto", cujo *recall* aumenta de $0,43$ para $0,52$. Embora a classe "baixo" apresente uma leve redução em seu *recall*, o desempenho global do modelo torna-se mais homogêneo, refletido no aumento do F1-score médio macro, que evolui de $0,51$ para $0,54$.

Apesar de a acurácia global permanecer praticamente inalterada ($0,60$ para $0,59$), os ganhos observados nas métricas de média macro e na sensibilidade às classes menos representadas indicam que o balanceamento híbrido contribui para uma melhor capacidade de generalização do modelo.

5 Conclusões e Sugestões para Trabalhos Futuros

Objetivou-se com este trabalho desenvolver um sensor virtual para classificação em tempo real do teor de sílica no concentrado de uma planta de flotação reversa de minério de ferro, utilizando técnicas de aprendizado de máquina em um banco de dados históricos do processo.

A aplicação do algoritmo *Random Forest* para as classificações binária e multiclasse evidenciou os desafios do uso de dados desbalanceados. A versão inicial do modelo para duas classes, apesar de uma acurácia global de 80%, demonstrou um *recall* de apenas 10% para a classe "Alta". Na prática, tal resultado indica que o modelo falharia em detectar a maioria dos eventos em que o teor de sílica está acima do especificado. Em um contexto de controle de processos, esse cenário é considerado grave, pois um falso negativo levaria a não atuação ou a atuação insuficiente do sistema de controle avançado. Nesse sentido, uma consequência direta seria a subdosagem de amina, o que prejudicaria a recuperação do minério de ferro.

A introdução da técnica híbrida de balanceamento representou um avanço significativo, equilibrando o *trade-off* entre precisão e revocação. A melhoria expressiva no *recall* da classe "Alta", com uma acurácia ainda robusta de 76%, produziu um modelo mais confiável e seguro para aplicação industrial. Para a classificação em três faixas, a técnica também promoveu uma distribuição mais homogênea do desempenho entre as classes, aumentando a sensibilidade do modelo às condições intermediárias e críticas.

Os resultados validam a hipótese de que um sensor virtual de classificação pode ser uma ferramenta valiosa para reduzir a dependência das análises laboratoriais periódicas. A saída categórica e em tempo real deste sensor pode ser diretamente integrada à camada de controle avançado existente na planta. Por exemplo, uma classificação acima do teor especificado poderia acionar automaticamente um incremento na dosagem de amina ou um ajuste no SP de nível das células para aumentar a velocidade de transbordo, ações que visam corrigir proativamente o desvio de qualidade. Dessa forma, o sensor virtual atua como um elemento de *feedforward*, antecipando correções e mitigando as perdas de eficiência.

Durante o desenvolvimento, identificou-se a limitação técnica referente ao uso das variáveis de velocidade e RGB, provenientes das câmeras instaladas nas células, devido a inconsistências de medição que demandam recalibração. Isso reforça que o sucesso de sensores virtuais depende, primordialmente, de uma instrumentação robusta e confiável. Essa restrição técnica ressalta a importância da manutenção dos ativos físicos como pré-requisito para a implementação de novas soluções de controle.

Ressalta-se que ambas as configurações do modelo — tanto a classificação binária quanto a multiclasse — mostraram-se aptas para serem testadas em ambiente operacional. No entanto, embora ambas as opções sejam utilizáveis, sua implementação nos sistemas de controle não deve ocorrer de forma simultânea, devendo-se optar pela configuração que melhor se adeque

aos objetivos de operação no momento. Dessa forma, o modelo pode integrar-se à camada de controle avançado, de forma a permitir ajustes ágeis na dosagem de reagentes e no nível das células.

Para trabalhos futuros, recomenda-se:

1. A exploração de outras técnicas de balanceamento e algoritmos de classificação mais robustos (como XGBoost ou Redes Neurais) para aumentar o desempenho;
2. A inclusão de outras variáveis de entrada, como características da espuma obtidas por visão computacional;
3. a realização de um estudo piloto de integração em tempo real, em que as previsões do sensor virtual alimentariam um controlador avançado, permitindo a validação prática de seus benefícios operacionais e econômicos;
4. A implementação de um algoritmo de regressão, visando obter a previsão do valor percentual de sílica no concentrado em tempo real.
5. Abordar o problema no contexto da detecção de anomalias.

Conclui-se, então, que o sensor virtual desenvolvido, adotando uma estratégia híbrida, é uma ferramenta viável para fornecer estimativas categóricas em tempo real do teor de sílica. Dessa forma, a solução apresenta potencial para aumentar a rentabilidade e a qualidade do produto final da Mina de Timbopeba, da Vale S.A..

Referências

AI, Mingxi; XIE, Yongfang; TANG, Zhaohui; ZHANG, Jin; GUI, Weihua. Deep learning feature-based setpoint generation and optimal control for flotation processes. *Information Sciences*, Elsevier, v. 578, p. 644–658, 2021. DOI: 10.1016/j.ins.2021.07.060. Disponível em: https://doi.org/10.1016/j.ins.2021.07.060. Citado 1 vez na página 11.

AVEVA. *AVEVA PI System*. Disponível em: https://www.aveva.com/pt-br/products/aveva-pi-system/. Acesso em: 6 fev. 2026. Citado 0 vez na página 29.

CAMPOS, Mario Cesar Massa de. *Controles típicos de equipamentos e processos industriais*. São Paulo: Blucher, 2010. Citado 1 vez na página 10.

CAMPOS, Mario Cesar Massa de; GOMES, Marcos Vinícius de Carvalho; PEREZ, José Manuel Gonzalez Tubio. *Controle avançado e otimização na indústria do petróleo*. Rio de Janeiro: Interciência, 2013. Citado 1 vez na página 10.

CHAVES, Arthur Pinto; LEAL FILHO, Laurindo de Salles; BRAGA, Paulo Fernando Almeida. Flotação. In: TRATAMENTO de Minérios. 6. ed. Rio de Janeiro: CETEM/MCTIC, 2018. cap. 10. Disponível em: http://mineralis.cetem.gov.br/handle/cetem/2181. Citado 8 vezes nas páginas 9, 13, 14.

DINIZ, Diego Rafael Monteiro. *Desenvolvimento de um soft sensor para inferência de eficiência energética de moinho de bolas em circuito fechado a úmido*. 2020. Dissertação de Mestrado – Universidade Federal de Ouro Preto, Ouro Preto. Disponível em: http://www.repositorio.ufop.br/jspui/handle/123456789/14049. Citado 4 vezes nas páginas 16, 18, 19.

ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. *Introdução a data science: algoritmos de machine learning e métodos de análise*. São Paulo: Casa do Código, 2020. Citado 1 vez na página 22.

FERNÁNDEZ, Alberto; GARCÍA, Salvador; PRATI, Mikel; HERRERA, Bartosz. *Learning from Imbalanced Data Sets*. Springer, 2018. Citado 7 vezes nas páginas 23–25.

GÉRON, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2022. Citado 6 vezes nas páginas 22, 23.

GUEDES, Érick Victor de Oliveira. *Aplicação de soft sensor baseado em redes neurais artificiais e random forest para predição em tempo real do teor de ferro no concentrado da flotação de minério de ferro*. 2020. Dissertação de Mestrado – Instituto Tecnológico da Vale, Ouro Preto. Disponível em: https://repositorio.itv.org/items/ddfdecf2-0f1b-4c8a-a3a3-19f1b928c3ac. Citado 1 vez na página 28.

JIANG, Xiaoping; ZHAO, Huilin; LIU, Junwei. Classification of Mineral Foam Flotation Conditions Based on Multi-Modality Image Fusion. *Applied Sciences*, v. 13, n. 6, 2023. DOI: 10.3390/app13063512. Disponível em: https://doi.org/10.3390/app13063512. Citado 2 vezes na página 27.

KADLEC, Petr; GABRYS, Bogdan; STRANDT, Sibylle. Data-driven Soft Sensors in the process industry. *Computers Chemical Engineering*, v. 33, n. 4, p. 795–814, 2009. ISSN 0098-1354. DOI: 10.1016/j.compchemeng.2008.12.012. Disponível em: https://doi.org/10.1016/j.compchemeng.2008.12.012. Citado 10 vezes nas páginas 16–19.

LIAO, Yinfei; LIU, Jiongtian; WANG, Yongtian; CAO, Yijun. Simulating a fuzzy level controller for flotation columns. *Mining Science and Technology (China)*, v. 21, n. 6, p. 815–818, 2011. ISSN 1674-5264. DOI: 10.1016/j.mstc.2011.05.038. Disponível em: https://doi.org/10.1016/j.mstc.2011.05.038. Citado 1 vez na página 9.

LIPTÁK, Béla G.; VENCZEL, Kriszta. *Measurement and Safety*. 5. ed.: CRC Press, 2016. v. 1. DOI: 10.4324/9781315370330. Disponível em: https://doi.org/10.4324/9781315370330. Citado 7 vezes nas páginas 16–19.

LIU, Bingyu; LIU, Bojian; GAO, Xianwen; ZHANG, Dingsen; HAO, Dezhi; LI, Xinyang. A soft sensor based on case-based reasoning for iron ores flotation. *Ironmaking & Steelmaking*, v. 47, n. 2, p. 150–158, 2020. DOI: 10.1080/03019233.2018.1497760. Disponível em: https://doi.org/10.1080/03019233.2018.1497760. Citado 1 vez na página 28.

LIU, Yuxi Hayden. *Principal Component Analysis*. 2. ed. UK: Springer, 2002. Citado 4 vezes nas páginas 25, 26.

LIU, Yuxi Hayden. *Python Machine Learning By Example: Unlock machine learning best practices with real-world use cases*. 4. ed. Birmingham, UK: Packt Publishing, 2024. Citado 3 vezes nas páginas 19–22.

LUZ, Adão Benvindo da; SAMPAIO, João Alves; FRANÇA, Silvia Cristina Alves. *Tratamento de minérios*. 5. ed. Rio de Janeiro: CETEM/MCT, 2010. Disponível em: http://mineralis.cetem.gov.br/handle/cetem/476. Citado 2 vezes na página 9.

PURAL, Yusuf Enes. Developing a data-driven soft sensor to predict silicate impurity in iron ore flotation concentrate. *Physicochemical Problems of Mineral Processing*, v. 59, n. 5, p. 169823, 2023. ISSN 1643-1049. DOI: 10.37190/ppmp/169823. Disponível em: https://doi.org/10.37190/ppmp/169823. Citado 1 vez na página 28.

RAMOS, Kerollan; FRADE, Altieres; SANTOS, Iranildes; PINTO, Thomás. Interpretable Prediction of Silica Content in Iron Ore Flotation Using Machine Learning. *IFAC*, Elsevier, v. 59, n. 32, p. 132–137, 2025. DOI: 10.1016/j.ifacol.2025.12.409. Disponível em: https://doi.org/10.1016/j.ifacol.2025.12.409. Citado 2 vezes na página 26.

SILVA, F.T.D; GALÉRY, R. Classicador Qualitativo de Teor de Sílica no Concentrado de Flotação Reversa de Minério de Ferro por Tamanho de Bolhas. *VIII Meeting of the Southern Hemisphere on Mineral Technology*, 2013. Disponível em: https://publicacoes.entmme.org/filebase/2013/2348%20-%20SILVA,%20F.T.D.-%20CLASSIFICADOR%20QUALITATIVO%20DE%20TEOR%20DE%20S%C3%8DLICA%20NO%20CONCENTRADO%20DE%20FLOTA%C3%87%C3%83O%20REVERSA%20DE%20MIN%C3%89RIO%20DE%20FERRO%20POR%20TAMANHO%20DE%20BOLHAS.PDF. Citado 2 vezes na página 27.

STRUTZEL, Flavio. *Controle IHMPC de um processo industrial de hidrotratamento de diesel*. 2014. Dissertação de Mestrado – Universidade de São Paulo - Programa de Engenharia Química, São Paulo. DOI: 10.11606/D.3.2014.tde-24062014-102335. Disponível em: https://doi.org/10.11606/D.3.2014.tde-24062014-102335. Citado 1 vez na página 10.

VALE. *Manutenção do Sistema Especialista da Flotação - Timbopeba*. 2021. Documento interno da Empresa VALE S.A. Acesso restrito. Citado 1 vez nas páginas 14, 16.

ZHANG, Dingsen; GAO, Xianwen. A digital twin dosing system for iron reverse flotation. *Journal of Manufacturing Systems*, v. 63, p. 238–249, 2022. ISSN 0278-6125. DOI: 10.1016/j.jmsy.2022.03.006. Disponível em: https://doi.org/10.1016/j.jmsy.2022.03.006. Citado 1 vez na página 28.

ZHANG, Dingsen; GAO, Xianwen. Soft sensor of flotation froth grade classification based on hybrid deep neural network. *International Journal of Production Research*, Taylor & Francis, v. 59, n. 16, p. 4794–4810, 2021. DOI: 10.1080/00207543.2021.1894366. Disponível em: https://doi.org/10.1080/00207543.2021.1894366. Citado 3 vezes nas páginas 11, 27.

ZHANG, Dingsen; GAO, Xianwen; QI, Wenhai. Soft sensor of iron tailings grade based on froth image features for reverse flotation. *Transactions of the Institute of Measurement and Control*, v. 44, n. 15, p. 2928–2940, 2022. DOI: 10.1177/01423312221096450. Disponível em: https://doi.org/10.1177/01423312221096450. Citado 1 vez na página 28.