Machine Learning no processamento do nióbio: aplicação da regressão *Random Forest* para análise de dados experimentais

João Gabriel Niquini Cordeiro [1]*
Cássia Ribeiro Souza [1]
Janúbia Cristina Bragança da Silva Amaral [1]
Sônia Denise Ferreira Rocha [1]
Pedro Henrique Alves Campos [1]

Resumo

Com os avanços tecnológicos na área computacional, o *Machine Learning (ML),* ou em português "Aprendizado de Máquina", emergiu como ferramenta poderosa para aplicações na ciência de dados, destacando-se a sua capacidade de fazer previsões baseadas em análises avançadas de dados por meio da identificação de padrões e relações implícitas. Nesse sentido, o objetivo deste trabalho é avaliar o seu desempenho de previsão no campo experimental a partir de dados hidrometalúrgicos referentes ao nióbio. Nesta perspectiva, foram feitos dois estudos de casos distintos: o primeiro a respeito do equilíbrio do ácido nióbico em soluções aquosas, e o segundo envolvendo a técnica de *sub-molten salt* e lixiviação alcalina de concentrados de pirocloro. O ML foi aplicado através do regressor *Random Forest* (RF). Os resultados mostram que para os dois estudos de caso, o erro gerado pelo modelo de RF foi inferior ao previsto pelas metodologias tradicionais, o que corrobora o benefício do uso de técnicas de ML em análises experimentais.

Palavras-chave: *Machine Learning*; Dados experimentais; *Random Forest*; Nióbio.

Machine Learning in niobium processing: application of *Random Forest* regression for experimental data analysis

Abstract

With technological advancements in computational fields, Machine Learning (ML) has emerged as a powerful tool for data science applications, particularly for its ability to make predictions based on advanced data analyses through the identification of patterns and implicit relationships. This study aims to evaluate the predictive performance of ML in the experimental context using hydrometallurgical data related to niobium. Two distinct case studies were conducted: the first on niobic acid equilibrium in aqueous solutions, and the second on the sub-molten salt method and leaching of pyrochlore concentrates. ML was applied using the *Random Forest* (RF) regressor. The results demonstrate that, in both case studies, the error generated by the RF model was lower than that predicted by traditional methodologies, supporting the advantages of ML techniques in experimental analyses.

Keywords: Machine Learning; Experimental data; *Random Forest*; Niobium.

1 Introdução

Machine learning (ML) é um subcampo da inteligência artificial que se concentra no desenvolvimento de algoritmos capazes de aprender e melhorar seu desempenho com base em dados. Ao invés de seguir explicitamente regras programadas, esses algoritmos identificam padrões e fazem previsões com base em informações previamente observadas [1]. Um dos vários campos nos quais o ML pode ser utilizado como ferramenta é o da análise de dados de experimentos, permitindo uma compreensão mais profunda dos fatores que influenciam os resultados ao identificar padrões complexos e interações entre variáveis que poderiam passar despercebidas em análises tradicionais [2-5].

[1]*Departamento de Engenharia de Minas – DEMIN, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, MG, Brasil.*

**Autor correspondente: jgncordeiroufmg@gmail.com*

E-mails: cassiaribeirosouza@yahoo.com.br; djannamaral@yahoo.com.br; sdrocha@demin.ufmg.br; pedrocampos@demin.ufmg.br

Algoritmos de aprendizado supervisionado (regressão e classificação) podem ser usados para prever os resultados com base em variáveis independentes, enquanto métodos não supervisionados (agrupamento) ajudam a agrupar dados com características semelhantes. Entre as técnicas de regressão de ML, se destacam o *Random Forest* (RF) [6,7], *Principal Component Regression* (PCR) e *Partial Least Squares* (PLS) [8], *Support Vector Machine* (SVM) [9] e *Artificial Neural Networks* (ANN) [10,11]. Algumas técnicas, entre elas ANN e RF, permitem a modelagem de interações não-lineares entre fatores, facilitando a quantificação da influência de cada variável nos resultados experimentais. Isso fornece uma visão mais robusta e precisa, auxiliando na tomada de decisões e na otimização de processos experimentais.

Na área mineral, o uso do ML pode ser dividido em três classificações: detecção/diagnóstico de erros, *machine vision* (MV) e modelamento com base em dados [2]. Em termos desta última categoria, os modelos (regressores ou classificadores) são aplicados como "sensores" para as variáveis infrequentes ou difíceis de serem medidas dos processos minerais. Ao utilizar da capacidade preditiva da máquina no lugar da complexidade experimental, o pesquisador ganha novas informações a respeito do estudo sem alocar muitos recursos para isso. A literatura reporta aplicações de ML no processamento mineral na identificação de parâmetros de circuitos de moagem [3], o modelamento de hidrociclones [4] e na manutenção do desempenho da flotação através de parâmetros de processo [5].

Na aplicação referente à moagem [3], estimar a vida útil dos equipamentos do circuito é crucial para facilitar decisões de manutenção preventiva do sistema. Usando o ANN foi possível estabelecer uma estratégia para determinar essa vida útil sem interrupção da operação de moagem. As variáveis de interesse foram a espessura e a vida restante do revestimento do moinho, enquanto os parâmetros do processo consistiram no tipo do minério, vazão de alimentação do minério (ton/h), potência (kW), velocidade angular (% da velocidade crítica), torque (% do torque máximo), vazão de água (m^3/h), energia de moagem (kWh/tonelada) e carregamento (toneladas). Os resultados se mostraram positivos, com uma precisão de 90% e a performance da ANN foi consistente tanto para os dados de treino quanto para os de teste. Assim, os ganhos podem ser resumidos no fato de que não é necessária a paralisação da operação de moagem nem a utilização de equipamentos sofisticados para obtenção dos dados necessários para encontrar a vida útil restante do revestimento do moinho.

No caso do modelamento dos hidrociclones [4], a perspectiva era de definir a eficiência do hidrociclone, por meio da determinação de duas importantes variáveis de saída: os fluxos (*Underflow* e *Overflow*) e o tamanho de corte corrigido (d_{50c}), através da criação de um modelo de ANN. As variáveis de entrada foram a queda de pressão na entrada (psi), a concentração de sólidos, o diâmetro do apex(mm) e o diâmetro do vórtex (mm), as quais foram medidas em um hidrociclone em escala de laboratório. Os resultados proveram um R^2 de 97,7% para o d_{50c} e de 98,9%/99,4% para o *Overflow*/*Underflow*, respectivamente, o que é promissor para o uso de ML para o controle automático da eficiência de hidrociclones.

A aparência da espuma pode ser considerada um indicativo forte da performance da flotação [5].

Com base nisso, uma ANN de três camadas foi criada, com as seguintes variáveis de saída: teor do concentrado de cobre (G_{Cu}), recuperação de cobre (R_{Cu}), recuperação de massa (R_m) e recuperação de água (R_w), as quais são determinantes para a avaliação do desempenho da operação. Nesse caso, as variáveis de entrada foram obtidas em conjunto com algoritmos de MV (porque dependem de fontes visuais de dados), sendo elas: tamanho da bolha, velocidade da espuma, cor da espuma e estabilidade da espuma. Foi encontrada uma forte correlação entre os fatores visuais de entrada (particularmente o tamanho da bolha e a velocidade da espuma) e os fatores de performance (R^2 de 90% para o R_{Cu} e de 97% para o restante das variáveis), o que é muito relevante para fins de controle automático do processo.

Foram propostos neste trabalho a aplicação de métodos de ML em dois estudos de caso de experimentos ligados ao processamento de minério de nióbio, um referente ao equilíbrio do ácido nióbico em solução alcalina e o outro referente à recuperação de nióbio após processo *sub-molten-salt* e lixiviação alcalina de impurezas contidas em finos de concentrado de pirocloro. Não foram identificados trabalhos na literatura que tratem especificamente do tema abordado sob a mesma perspectiva adotada neste estudo. Essa ausência de publicações similares evidencia o caráter inovador do trabalho, indicando uma lacuna ainda não explorada pela comunidade científica.

No caso do estudo do equilíbrio do ácido nióbico em soluções aquosas, o método de *Machine Learning* foi comparado com o método espectrométrico baseado na lei de Lambert-Beer, um método de análise mais barato e simples, mas que pode apresentar desvios analíticos quando comparado à métodos mais exatos como o ICP-OES. Os modelos *Decision Trees*, *Elastic Net*, *Support Vector Machine* (SVM), *Gradient Boosting* (GB) e *Random Forest* (RF) foram testados para identificar qual produz melhor resultado no contexto desses dados, sendo que o RF apresentou o melhor resultado.

No segundo estudo, foram consideradas duas variáveis de resposta: a recuperação de nióbio e a remoção de sílica após a lixiviação do material gerado pelo método do *sub-molten salt*. As variáveis de entrada foram definidas previamente por meio de um Planejamento Composto Central (PCC), e a partir desse conjunto de dados, foram ajustados modelos preditivos utilizando *Random Forest* e regressão linear, com o objetivo de comparar seus desempenhos.

O RF utiliza como "classificador base" as árvores de decisão para determinar o valor resultante com base nos parâmetros entregues, disso decorre o nome de "Floresta". Individualmente, cada árvore de decisão se comporta em termos de raízes, galhos e folhas, onde as raízes são as

variáveis de entrada, as folhas as variáveis de saída (as classes) e os galhos, os valores das variáveis de entrada [12]. Aumentando o escopo do problema, o *Random Forest* é um algoritmo de "*ensemble*" (agrupamento), no qual ela agrupa diferentes classificadores, nesse caso as árvores de decisão, e usa alguma técnica de agregação como por exemplo "voto da maioria", de modo que os classificadores cheguem, em sua maioria, na mesma conclusão [1].

2 Objetivos

O objetivo deste estudo é a aplicação de métodos de aprendizado de máquina na análise de resultados de experimentos relacionados à indústria mineral de nióbio, com o objetivo de encontrar melhores modelos de predição. Para isso, são utilizados dois bancos de dados de experimentos diferentes, a saber:

- Dados de solubilidade do composto de nióbio – ácido nióbico -em solução de hidróxido de potássio. O objetivo é comparar a modelagem preditiva através de técnicas de ML com resultados analíticos obtidos utilizando o modelo espectrométrico baseado na lei de Lambert-Beer para quantificação do nióbio aquoso em solução;
- Dados resultantes de um planejamento de experimentos, no contexto da análise da recuperação de nióbio e remoção de sílica de concentrado de pirocloro após o método de *sub-molten salt* com NaOH e lixiviação alcalina. Nesse caso, o foco também é comparativo, mas o intuito é comparar a regressão linear simples com os resultados obtidos através de um modelo de *Random Forest*.

3 Metodologia

Nesta seção são apresentadas as variáveis de cada um dos bancos de dados, como elas foram medidas e qual informação elas representam. Também serão mostrados a técnica e os parâmetros utilizados para a modelagem do algoritmo.

3.1 Experimentos do estudo do equilíbrio de ácido nióbico

O estudo de equilíbrio utilizou como amostra o ácido nióbio comercial HY-340, no qual a massa do sólido foi colocada em 10 mL de soluções de 0,10 M e 0,25 M de hidróxido de potássio em recipientes de polietileno. Foram mantidas em agitação contínua de 200 RPM a 25 °C (incubadora *shaker* SOLAB SL-222) para se alcançar a condição de equilíbrio, indicada pela concentração de nióbio medida como uma função de tempo. As amostras foram filtradas utilizando um filtro seringa 0,45 µm (MILLIPORE MILLEX-HV PVDF 0,45 µm), e a medida do pH e a quantificação do Nb presente em solução foram feitos após a filtragem.

A concentração de nióbio solubilizado foi determinada pelos métodos ICP-OES e o UV-visível e, considerando o limite de detecção do ICP-OES (Perkin Elmer, model Optima 7300DV) de 0,005 mol/L, os valores obtidos por esse método foram considerados como o padrão de performance e exatidão no âmbito dessa análise.

O segundo método analítico foi baseado no trabalho de Deblond et al. [13], onde cada amostra foi diluída em 4 mol/L de hidróxido de sódio e reagidas por 10 minutos antes da medição da absorbância em espectrômetro UV-Visível em comprimento de onda de 246,5 nm, caminho ótico de 1 cm. A concentração total de nióbio se baseia na presença de dois complexos solúveis: $Nb_6O_{19}^{-8}$ e $HNb_6O_{19}^{-7}$. Foram utilizados os coeficientes de absortividade molar de 15900±600 L/mol.cm para a espécie $Nb_6O_{19}^{-8}$ e 14300±400 L/mol.cm para a espécie $HNb_6O_{19}^{-7}$. O espectro UV foi medido em um espectrofotômetro BEL-M51, com a amostra colocada em cubetas de quartzo. A equação da lei de Lambert-Beer foi utilizada para determinar a concentração de Nb em solução, através da Equação 1:

$$A = \varepsilon * c * l \quad (1)$$

onde:
- ε: coeficiente de absortividade molar;
- c: concentração da molécula no meio;
- l: caminho óptico;
- A: absorbância.

3.1.1 Variáveis do banco de dados para modelagem pelo *Random Forest*

Pela sua precisão e exatidão superior, o método ICP-EOS foi tomado como o valor da concentração de nióbio solubilizado mais exato, sendo o UV-Vis o método mais prático de ser efetuado, havendo diferenças metodológicas de obtenção dos resultados entre eles.

A partir disso, as variáveis do banco de dados serão os fatores presentes desde a medição da massa da amostra até os valores lidos no espectrofotômetro BEL-M51 que variam entre cada uma das corridas experimentais. As variáveis base presentes na lei de Lambert-Beer (ε e **l**) são desconsideradas nessa modelagem.

Assim sendo, as variáveis utilizadas no modelo foram o tempo (em horas), a quantidade de amostra do ácido nióbico (em gramas), a concentração de KOH (em mol/L) e a absorbância medida em cada corrida experimental do UV-vis (adimensional), conforme Tabela 1. Ademais, é válido ressaltar que os valores de absorbância utilizados foram a média de três valores, dado que para cada experimento, a absorbância foi medida três vezes para aquela condição experimental.

Por questões de confidencialidade, o banco de dados não pode ser disponibilizado. Entretanto, na Tabela 2 são

apresentados os parâmetros estatísticos mais importantes de cada variável no primeiro estudo de caso.

3.1.2 Aplicação do *Machine Learning*

Inicialmente, os modelos *Decision Trees*, *Elastic Net*, *Support Vector Machine* (SVM), *Gradient Boosting* (GB) e *Random Forest* (RF) foram testados para prever a concentração de nióbio solubilizado (variável dependente) a partir das variáveis independentes da Tabela 1. O método RF gerou resultados mais precisos, sendo escolhido para o trabalho, conforme pode ser observado pela Tabela 3, que evidencia a média e o desvio padrão (DP) de R^2 e RMSE para todos os métodos.

Com o método de análise estabelecido, é preciso dividir o banco de dados entre treino e teste, de modo a evitar o problema conhecido como *Overfit*. Esse é um problema também conhecido como sobre-ajuste, que ocorre quando os modelos de análise de dados são criados a partir do banco de dados completo e não de uma parte dele, o que torna o modelo resultante totalmente enviesado.

Em modelos formados a partir do conjunto completo de dados apresentados, se novos dados forem inseridos no banco, o modelo não é capaz de entregar um valor da variável de resposta que seja condizente com a realidade. Nas situações em que o conjunto de valores é devidamente subdivido entre teste e treino, o modelo explica mais naturalmente a variância inerente dos dados e apresentará um valor de resposta muito mais próximo do real, dependendo das condições nas quais ele foi desenvolvido (método escolhido, quantidade de dados, qualidade dos dados).

No presente trabalho, essa divisão foi feita através da criação de dez *folds*, que repartem de maneira aleatória o banco de dados em dados de teste e treino múltiplas vezes e criam um modelo para cada uma dessas tentativas. Dessa forma, o algoritmo consegue definir qual desses modelos de RF hipotéticos é melhor para explicar a variância presente nos dados e, assim, entregar o modelo mais preciso.

Para a construção e avaliação do modelo de RF, foi utilizada a biblioteca *Scikit-Learn*, amplamente adotada na linguagem Python por sua interface simples e eficiente para algoritmos supervisionados. A otimização dos hiper parâmetros do modelo de *Random Forest Regressor* foi realizada por meio da técnica de *Grid Search*, que consiste em testar todas as combinações possíveis de um conjunto pré-definido de valores para cada parâmetro, utilizando validação cruzada para identificar a configuração que proporciona o melhor desempenho segundo uma métrica estabelecida.

Os melhores resultados foram obtidos com os seguintes valores: *n_estimators* = 300 (número de árvores na floresta), *max_depth* = 20 (profundidade máxima de cada árvore), *max_features* = log2 (número de variáveis consideradas em cada divisão) e *min_samples_split* = 2 (mínimo de amostras para dividir um nó).

Tabela 1. Fragmento do banco de dados com as variáveis utilizadas

Tempo (h)	Nb_2O_5,nH_2O (g)	Concentração KOH (mol/L)	Média ABS
0,5	0,1901	0,25	0,032000
2,0	0,2000	0,25	0,014667
2,0	0,2000	0,25	0,014667
4,0	0,1987	0,25	0,045000
8,0	0,1973	0,25	0,020000
16,0	0,2020	0,25	0,081000

Tabela 2. Estatísticas das variáveis do primeiro banco de dados

Parâmetros	Tempo (h)	Nb_2O_5,nH_2O (g)	Concentração KOH (mol/L)	Média ABS
Tamanho da Amostra	45	45	45	45
Média	216,12	0,199	0,167	0,164
Desvio Padrão	193,93	0,002	0,075	0,153
Mediana	175	0,200	0,1	0,088
Mínimo	0,5	0,190	0,1	0,007
Máximo	744	0,202	0,25	0,483

Tabela 3. Média e desvio padrão de RMSE e R^2 para cada método

Modelos	Média do RMSE	DP do RMSE	Média do R^2	DP do R^2
Decision Tree	4,52	3,26	0,87	0,12
Elastic Net	15,64	4,71	0,44	0,11
SVM	20,92	11,51	0,01	0,15
GB	4,14	2,98	0,93	0,09
RF	4,18	2,06	0,96	0,02

3.2 Experimentos do estudo da recuperação de nióbio e remoção de sílica – *Sub-molten salt* e lixiviação alcalina

Nesse estudo, finos de concentrado de pirocloro são submetidos ao método de *sub-moltem salt* com NaOH para promover a formação de silicatos solúveis e, posteriormente, a de lixiviação com água destilada [14-16]. Na primeira etapa do experimento, o NaOH em micro-pérolas ("pellets") e a amostra são pesados para atender uma devida proporção. A mistura é homogeneizada em um moinho de bancada e levada à mufla em tempos diversos.

Após o resfriamento a temperatura ambiente e em dessecador, os sólidos são lixiviados com água destilada por tempo determinado. Os sólidos são secos a 110 °C por duas horas, pesados e analisados quimicamente. A recuperação de nióbio e da remoção de sílica são calculadas pelas Equações 2 e 3:

$$Recuperação\ de\ Nb(\%) = \frac{Teor\ final\ de\ Nb * massa\ da\ torta\ lixiviada}{Teor\ inicial\ de\ Nb * massa\ de\ amostra\ inicial} \quad (2)$$

$$Remoção\ de\ Si(\%) = 1 - \frac{Teor\ final\ de\ Si * massa\ da\ torta\ lixiviada}{Teor\ inicial\ de\ Si * massa\ de\ amostra\ inicial} \quad (3)$$

3.2.1 Variáveis do banco de dados para modelagem pelo *Random Forest*

Antes da execução das corridas experimentais, com base na literatura [14-16], foi realizado um Planejamento Composto Central (PCC) [17]. As variáveis selecionadas como pertinentes tanto para a recuperação de nióbio quanto para remoção de sílica foram: concentração de NaOH (g/Kg), tempo (min) e temperatura (°C) do tratamento térmico, tempo de lixiviação (min) e concentração de sólidos (%).

O PCC cria o conjunto de experimentos em intervalos fixos de valores para cada uma das variáveis (*e.g.* a concentração de sólidos varia em 15%, 30% e 45%), como pode ser visto na Tabela 4, o que o permite supor relações lineares e/ou polinomiais entre as variáveis do experimento.

Complementando a visualização apresentada na Tabela 4 do Planejamento Composto Central (PCC), a Tabela 5 exibe o intervalo de variação (*range*) de cada uma das variáveis de entrada do banco de dados. Ao todo, foram analisadas 43 amostras experimentais. Por motivos de confidencialidade, as variáveis de saída foram omitidas de ambas as tabelas.

3.2.2 Aplicação do *Machine Learning*

O algoritmo *Random Forest* (RF) foi selecionado após testes comparativos com outros métodos de aprendizado de máquina, visando sua comparação com a regressão linear simples, ambos ajustados pela mesma biblioteca em Python (*Scikit-learn*). Assim como no primeiro estudo de caso, a divisão dos dados em conjuntos de treino e teste foi adotada para mitigar o risco de *overfitting* — uma preocupação ainda mais relevante devido à quantidade limitada de dados amostrais.

Nessa etapa, tanto o RF quanto a regressão linear foram avaliados por meio de validação cruzada com subdivisão em *folds*, sendo então determinado o modelo com melhor ajuste para cada abordagem.

4 Resultados e discussão

4.1 Análise por ML do estudo de equilíbrio do ácido nióbico

Os gráficos da Figura 1a e 1b comparam, para a concentração de 0,10 mol/L de KOH e 0,25 mol/L de KOH, respectivamente, os valores da concentração de nióbio dissolvido em solução (%) em função do tempo por três métodos: o método considerado mais preciso (ICP-OES), a modelagem segundo o UV-Vis e o modelo de RF criado a partir dos dados.

Fica evidente que a curva dos dados do modelo de RF se aproxima muito mais ao padrão representado pela curva ICP do que a curva da metodologia UV suportada pela lei de Lambert-Beer. Para explicitar ainda mais essa diferença com valores, calculou-se a média e o desvio

Tabela 4. Fragmento do segundo banco de dados com as variáveis

Corrida	Temperatura do tratamento térmico (°C)	NaOH (g/Kg)	Tempo do tratamento térmico (min)	Concentração de Sólidos (%)	Tempo de Lixiviação (min)
1	400,0	300,0	30,0	15,0	30,0
2	700,0	300,0	30,0	45,0	30,0
3	400,0	300,0	60,0	15,0	60,0
4	700,0	300,0	60,0	15,0	60,0
5	400,0	600,0	30,0	30,0	30,0

Tabela 5. Valores das variáveis de entrada do segundo banco de dados

Temperatura do tratamento térmico (°C)	193	400	550	700	907
NaOH (g/Kg)	93	300	450	600	807
Tempo do tratamento térmico (min)	9	30	45	60	81
Concentração de Sólidos (%)	5	15	22,5	30	40
Tempo de Lixiviação (min)	9	30	45	60	81

padrão do coeficiente R^2 e do RMSE (raiz do erro quadrático médio) para as amostras de teste em cada *fold*, como pode ser observado na Tabela 6, tanto para Lambert-Beer (LB) quanto para o modelo de *Random Forest* (RF).

No entanto, é necessário evidenciar que, em termos de poder de predição, o modelo de RF fica limitado se ele é treinado apenas com uma das concentrações, tal como foi feito na análise acima. Caso novas corridas experimentais com concentrações diferentes das treinadas forem feitas, o modelo não preverá com muita confiabilidade resultados a partir delas.

Tabela 6. Média e desvio padrão de R^2 e RMSE para cada concentração

Conc. 0.1 mol/L	Média	Desvio Padrão
R^2 (RF)	0,78	0,22
RMSE (RF)	1,48	0,66
R^2 (LB)	0,58	0,20
RMSE (LB)	3,55	2,32
Conc. 0,25 mol/L	-	-
R^2 (RF)	0,96	0,04
RMSE (RF)	4,43	2,77
R^2 (LB)	0,89	0,08
RMSE (LB)	5,34	3,08

Assim sendo, foi desenvolvido um terceiro modelo de RF utilizando ambas as concentrações do banco de dados, no intuito de aumentar esse poder de predição e para analisar, consequentemente, a variação no parâmetro do erro gerada a partir dessa mudança. Novamente, a Figura 2 representa a diferença entre os modelos para a concentração de 0,1 mol/L (à esquerda), e 0,25 mol/L (à direita).

Embora visualmente não haja diferença clara entre as Figuras 1a e 2a (e entre a 1b e 2b), é preciso recalcular a média de R^2 e de RMSE e analisar se o desempenho do RF continua superior ao uso do método espectrométrico-Lei de Lambert-Beer. Os resultados podem ser vistos na Tabela 7.

Os resultados dessa aplicação sugerem que a concentração de nióbio solubilizado pode ser estimada por RF a partir das variáveis de entrada, tempo (em horas), a quantidade de amostra do ácido nióbico (em gramas), a concentração de KOH (em mol/L) e a absorbância média obtida pelo UV-vis, sem a necessidade de fazer a medida por ICP e com erro menor do que o UV-Vis.

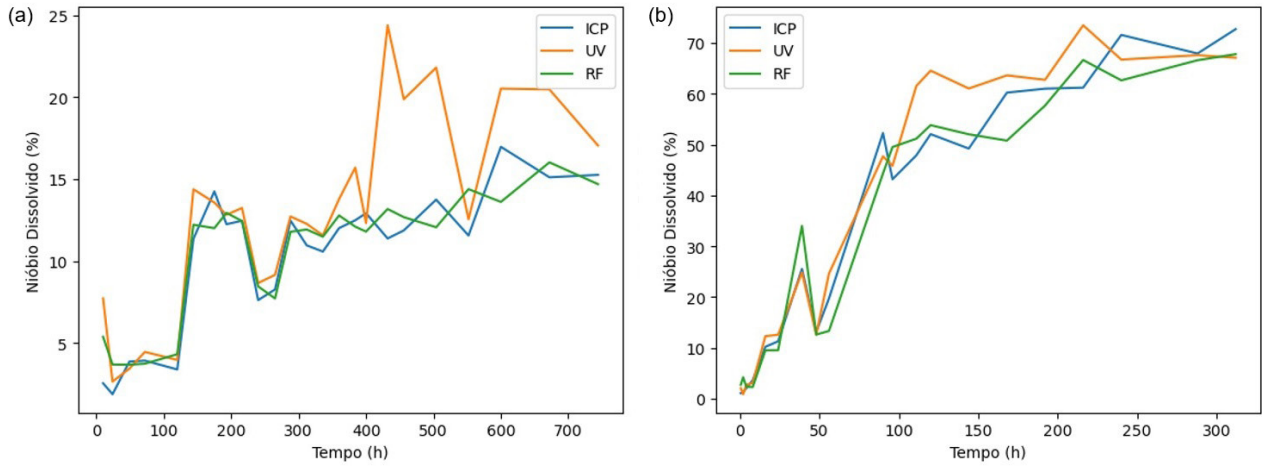

Figura 1. Comparação entre os valores previstos e experimentais da "concentração" de nióbio em 0.1 mol/L (a) e 0,25 mol/L (b) de KOH.

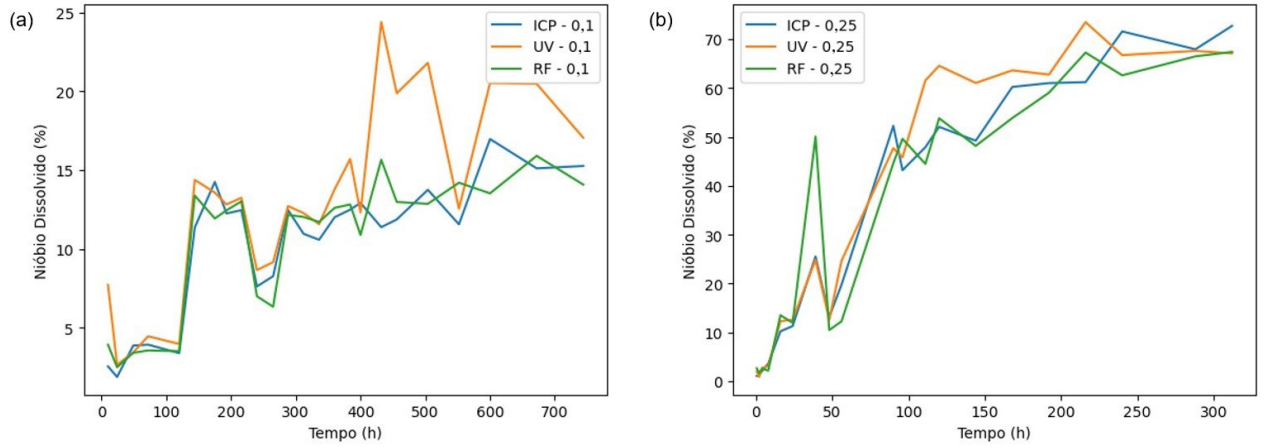

Figura 2. Comparação entre valores experimentais e modelo RF treinado em 0.1 mol/L (a), e 0.25 mol/L (b) de KOH.

Para complementar essa sugestão, foi efetuado a análise da importância (*feature importance*) de cada uma das variáveis de acordo com o RF, conforme Figura 3:

Pode ser observado que a absorbância obtida pela espectrometria é o fator predominante dentre as variáveis. Por conta disso, foi construído um gráfico (Figura 4) de absorbância por concentração de nióbio para analisar como está disposta a relação entre essas variáveis.

Em uma primeira observação, a Figura 4 sugere que para valores mais baixos de absorbância e de nióbio dissolvido, a relação entre essas variáveis é extremamente linear. No entanto, para concentrações mais elevadas, esse comportamento linear é perdido.

Apesar da expectativa inicial de que a concentração de nióbio dissolvido aumentasse continuamente com o tempo de contato com a solução alcalina, observou-se um comportamento instável em algumas condições experimentais, especialmente em concentrações mais elevadas de KOH.

Tabela 7. Média e desvio padrão de R^2 e RMSE para ambas concentrações

Random Forest	**Média**	**Desvio Padrão**
R^2	0,96	0,06
RMSE	3,28	1,78
Lambert-Beer	-	-
R^2	0,70	0,43
RMSE	4,45	2,42

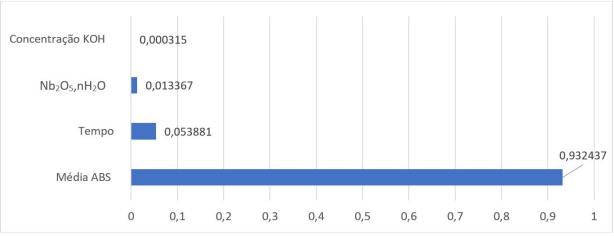

Figura 3. *Feature importance* das variáveis do estudo.

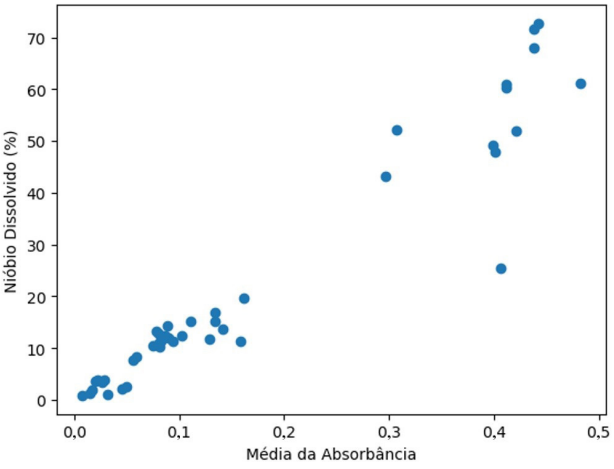

Figura 4. Relação entre a média da absorbância e a concentração de nióbio dissolvido.

Esse fenômeno pode ser explicado por processos de reprecipitação de espécies de nióbio, como polioxoniobatos, ou pela formação de camadas superficiais de hidróxidos na superfície, limitando a continuidade da dissolução. Esse aspecto será objeto de trabalhos futuros.

Além disso, a supersaturação local pode induzir a nucleação de fases sólidas secundárias, reduzindo temporariamente a concentração de Nb em solução. Esses efeitos são consistentes com observações anteriores feitas por Deblond et al. [13], que relataram comportamento similar em sistemas contendo Nb(V) sob condições alcalinas brandas.

4.2 Análises por ML do estudo de recuperação do nióbio e remoção da sílica dos finos do concentrado de pirocloro

Para evidenciar a diferença de desempenho entre a regressão linear simples e o algoritmo de *Random Forest* (RF), os gráficos da Figura 5 apresentam a relação entre os erros de ambos os modelos (Equação 4). Os modelos foram ajustados a partir de uma divisão entre conjuntos de treino e teste. Nos gráficos, os pontos vermelhos indicam os casos em que a regressão linear apresentou erro maior que o *Random Forest*, enquanto os pontos azuis representam os casos em que o *Random Forest* teve desempenho inferior.

$$Erro = \left(\left[valor\ experimental\right] - \left[valor\ do\ modelo\right]\right)^2 \quad (4)$$

A partir da análise da Figura 5b, observa-se que, em média, o modelo de *Random Forest* apresenta erros menores em comparação à regressão linear simples. Isso se deve à maior dispersão e magnitude dos erros no modelo linear, confirmando as expectativas iniciais sobre a maior flexibilidade do *Random Forest* — ainda que este seja mais suscetível ao sobreajuste. Por outro lado, a avaliação visual da Figura 5a não permite identificar com clareza qual dos modelos teve melhor desempenho. Para isso, a Tabela 8 apresenta a soma dos erros quadráticos de cada abordagem.

Vale destacar que, em casos específicos, tanto na Recuperação de Nióbio quanto na Remoção de Sílica, a regressão linear apresenta erros consideravelmente elevados, evidenciando limitações em sua capacidade de generalização.

Para entender melhor como os parâmetros de entrada afetam os resultados, foram feitos os gráficos de *feature importance* para a Recuperação de Nióbio e para a Remoção de Sílica. As Figuras 6a e 6b abaixo demonstram a importância das variáveis de entrada, de 0 até 1:

Pode-se observar dois fatores importantes a partir desta figura:

- Para os dois casos, o tempo de digestão, de lixiviação e a concentração de sólidos foram pouco pertinentes para a criação dos modelos;
- Para ambos os modelos, a concentração de hidróxido de sódio foi a variável mais importante.

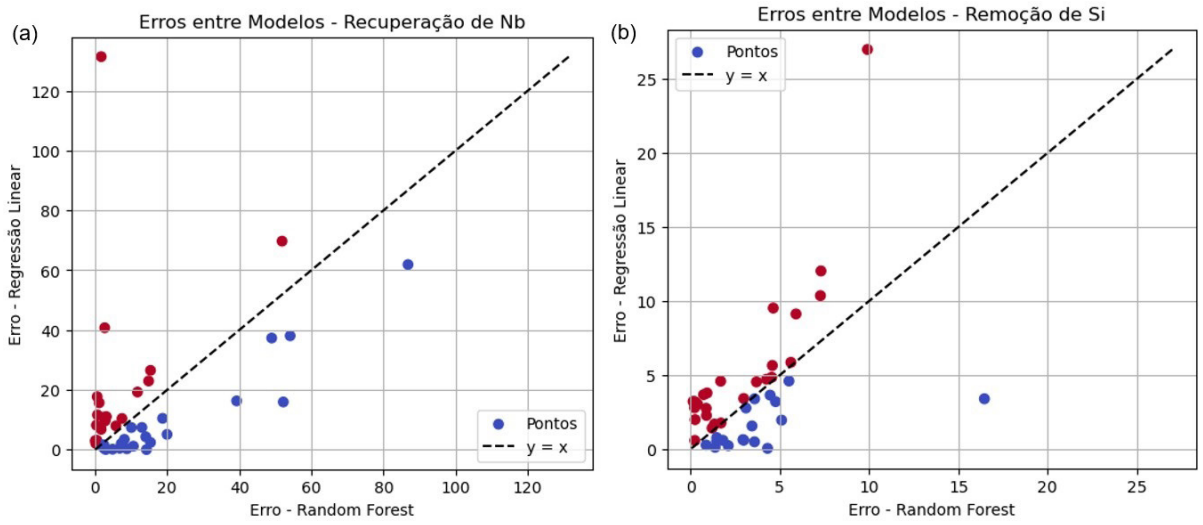

Figura 5. Comparação do erro utilizando Recuperação de Nióbio (a) e Remoção de Sílica (b) como variável resposta – divisão treino e teste.

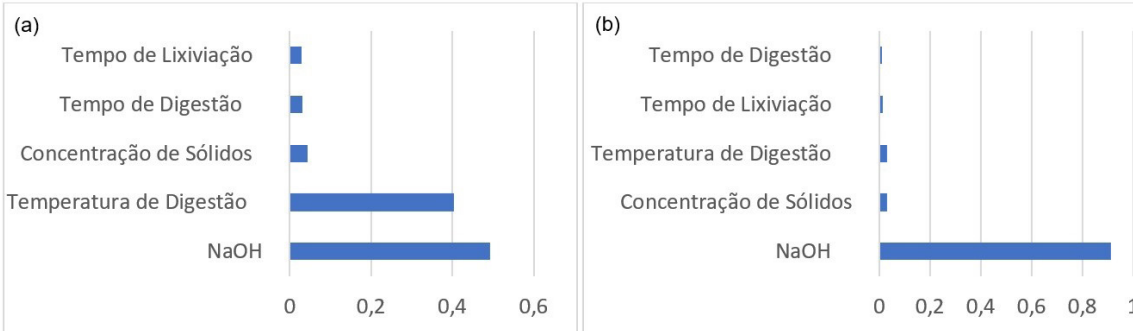

Figura 6. *Feature importance* para a recuperação de Nióbio (a) e a Remoção de Sílica (b).

Tabela 8. Média e desvio padrão de R^2 e RMSE para cada modelo

Recuperação de Nb	Média	Desvio Padrão
R^2 (RF)	0,62	0,22
RMSE (RF)	3,18	1,18
R^2 (RL)	0,52	0,37
RMSE (RL)	3,63	1,21
Remoção de Si	-	-
R^2 (RF)	0,88	0,04
RMSE (RF)	3,92	1,53
R^2 (RL)	0,79	0,17
RMSE (RL)	5,10	3,11

5 Conclusões

Com base nas hipóteses feitas inicialmente e nos valores encontrados nas análises do RF para ambos os bancos de dados, o ML se mostrou como uma ferramenta de uso viável e até mesmo superior em termos de predição quando comparado às técnicas tradicionais em cada caso. Apesar dos dois estudos estarem relacionados ao nióbio, o pretexto das problemáticas em termos de análise de dados eram bem diferentes e, mesmo assim, os modelos criados apresentaram resultados satisfatórios para ambas as predições.

No primeiro estudo, a curva de ABS em função da %Nb apresentou não linearidades, o que indica que o comportamento da variável resposta não segue uma relação completamente linear com a variável explicativa. Por isso, modelos não lineares tendem a se ajustar melhor aos dados, capturando essas nuances com maior precisão, como é o caso do *Random Forest*. No entanto, modelos não lineares mais simples, como regressões polinomiais de baixa ordem, podem ser suficientes para obter resultados próximos aos de modelos mais sofisticados. Como sugestão para trabalhos futuros, recomenda-se investigar com mais profundidade o desempenho comparativo entre modelos

não lineares e regressões polinomiais em diferentes intervalos de %Nb, avaliando sua robustez em cenários com maior complexidade ou variabilidade nos dados.

No segundo caso, considerando aplicações futuras, o modelo seria ajustado utilizando todo o conjunto de dados disponível, uma vez que não haveria mais a necessidade de separar dados para validação. Espera-se, portanto, que o erro final do modelo seja menor do que o erro observado durante a etapa de validação nos dados de treino. Essa abordagem visa maximizar o desempenho preditivo do modelo em cenários reais.

No contexto dos estudos representados e da literatura revisitada, é possível afirmar que o *Machine Learning* é uma alternativa a ser levada em conta quando comparada aos métodos usuais e que a tendência mundial é que sua utilização se torne cada vez mais comum.

Referências

1 Kubat M. An introduction to machine learning. 1st ed. Berlin: Springer; 2017. http://doi.org/10.1007/978-3-319-63913-0.

2 McCoy J, Auret L. Machine learning applications in minerals processing: a review. Minerals Engineering. 2019;132:95-109. http://doi.org/10.1016/j.mineng.2018.12.004.

3 Ahmadzadeh F, Lundberg J. Remaining useful life prediction of grinding mill liners using an artificial neural network. Minerals Engineering. 2013;53:1-8. http://doi.org/10.1016/j.mineng.2013.05.026.

4 Karimi M, Dehghani A, Nezamalhosseini A, Talebi S. Prediction of hydrocyclone performance using artificial neural networks. Journal of the Southern African Institute of Mining and Metallurgy. 2010 [acesso em 16 jan. 2025];110(5):207-212. Disponível em: http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S2225-62532010000500003

5 Jahedsaravani A, Marhaban MH, Massinaei M. Prediction of the metallurgical performances of a batch flotation system by image analysis and neural networks. Minerals Engineering. 2014;69:137-145. http://doi.org/10.1016/j.mineng.2014.08.003.

6 Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and regression trees. 1st ed. London: Chapman & Hall/CRC; 1984. http://doi.org/10.1201/9781315139470.

7 Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32. http://doi.org/10.1023/A:1010933404324.

8 Kettaneh N, Berglund A, Wold S. PCA and PLS with very large data sets. Computational Statistics & Data Analysis. 2005;48(1):69-85. http://doi.org/10.1016/j.csda.2003.11.027.

9 Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273-297. http://doi.org/10.1007/BF00994018.

10 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-444. http://doi.org/10.1038/nature14539.

11 Schmidhuber J. Deep learning in neural networks: an overview. Neural Networks. 2015;61:85-117. http://doi.org/10.1016/j.neunet.2014.09.003.

12 Mohanty S, Nanda R, Panigrahi P. Introduction to data science: a Python approach to concepts, techniques and applications. 1st ed. New York: Springer; 2019.

13 Deblonde GJP, Chagnes A, Bélair S, Cote G. Solubility of niobium(V) and tantalum(V) under mild alkaline conditions. Hydrometallurgy. 2015;156:99-106. http://doi.org/10.1016/j.hydromet.2015.05.015.

14 Shikika A, Sethurajan M, Muvundja F, Mugumaoderha MC, Gaydardzhiev S. A review on extractive metallurgy of tantalum and niobium. Hydrometallurgy. 2020;198:105496. http://doi.org/10.1016/j.hydromet.2020.105496.

15 Wang X, Zheng S, Xu H, Zhang Y. Leaching of niobium and tantalum from a low-grade ore using a KOH roast-water leach system. Hydrometallurgy. 2009;98(3–4):219-223. http://doi.org/10.1016/j.hydromet.2009.05.002.

16 Ayanda OS, Adekola FA. A review of niobium-tantalum separation in hydrometallurgy. Journal of Minerals & Materials Characterization & Engineering. 2011;10(3):245-256. http://doi.org/10.4236/jmmce.2011.103016.

17 Montgomery DC, Rumsby GL. Estatística aplicada e probabilidade para engenheiros. 6. ed. Rio de Janeiro: LTC; 2014.

Recebido em: 16 Jan. 2025
Aceito em: 29 Maio 2025

Editor responsável:
André Carlos Silva