UNIVERSIDADE FEDERAL DE UBERLÂNDIA

SRIRATNA SOUSA DE OLIVEIRA

USO DE MACHINE LEARNING NA MINERAÇÃO: Revisão de Literatura e Aplicação do Algoritmo *Random Forest* para Otimização da Recuperação Mássica Durante o Beneficiamento de Ferro

MONTE CARMELO

2022

SRIRATNA SOUSA DE OLIVEIRA

USO DE MACHINE LEARNING NA MINERAÇÃO: Revisão de Literatura e Aplicação do Algoritmo *Random Forest* para Otimização da Recuperação Mássica Durante o Beneficiamento de Ferro

Projeto de pesquisa apresentado como requisito ao desenvolvimento das atividades do Trabalho de Conclusão de Curso conforme a Norma para Elaboração de Trabalhos de Conclusão de Curso de Geologia.

Orientador: Prof. Dr. Emerson Rodrigo Almeida.

MONTE CARMELO

2022

UNIVERSIDADE FEDERAL DE UBERLÂNDIA CAMPUS MONTE CARMELO
INSTITUTO DE GEOGRAFIA – CURSO DE GRADUAÇÃO EM GEOLOGIA

SRIRATNA SOUSA DE OLIVEIRA

USO DE MACHINE LEARNING NA MINERAÇÃO: Revisão de Literatura e Aplicação do Algoritmo *Random Forest* para Otimização da Recuperação Mássica Durante o Beneficiamento de Ferro

Trabalho Final de Graduação para obtenção do grau de Bacharel em Geologia pela Universidade Federal de Uberlândia (UFU).

Banca Examinadora

_____

Professor Dr. Emerson Rodrigo Almeida (Orientador)
Universidade Federal do Uberlândia

_____

Professora Dra. Liliane Ibrahim
Universidade Federal do Uberlândia

_____

Msc. William Medina Leite
Universidade Federal dos Vales Jequitinhonha e Mucuri

“Não há saber mais ou menos: há saberes diferentes” - Pedagogia do Oprimido (1987)

Paulo Freire

AGRADECIMENTOS

Agradeço aos professores e funcionários do curso de geologia da UFU pela contribuição fundamental para minha formação como geólogo, ao professor Dr. Emerson pela orientação e contribuição para este trabalho e aos membros da banca Dra. Liliane e Msc. William Medina.

Gostaria de agradecer ao time de *Data Analytics* da Anglo American Brasil pela oportunidade e treinamento em métodos de ciência de dados em geologia, agradeço em especial ao grupo TS onde estou construindo e consolidando o aprendizado sobre *Machine Learning*.

Agradeço a Dra. Krishina pelos esclarecimentos de alguns conceitos e dicas de literatura e ao amigo e design Ivan pela ajuda com as figuras.

Por fim, gostaria de agradecer a minha família, cujo apoio e ajuda foram fundamentais durante toda minha graduação.

RESUMO

A mineração passa por mudanças importantes possibilitadas pelo desenvolvimento tecnológico exponencial das últimas décadas. Novos desafios ambientais e mudanças regulatórias fomentam a busca de ferramentas que possibilitem uma mineração mais eficaz e sustentável. Nesse contexto, a evolução dos métodos estatísticos e computacionais de inteligência artificial (IA) e *machine learning* (ML) são ferramentas importantes que permitem a análise massiva de dados de alta resolução e grande escala, permitindo aprimoramento de processos e aumento da produtividade. Uma aplicação pouco explorada é o uso de IA na redução de perdas durante o processo de beneficiamento. Na mineração de ferro, por exemplo, estima-se que as perdas nas usinas possam chegar a 20% e no Brasil de 20 a 40% do peso do total do minério lavrado é destinado para barragens de rejeito. Vários fatores estão relacionados a uma recuperação mássica inferior ao esperado, desde variáveis associadas ao beneficiamento na usina às associadas a composição mineralógica do sítio; no entanto, estabelecer a contribuição de cada uma dessas variáveis permanece um desafio. Assim, o objetivo deste trabalho é fazer uma revisão de literatura dos conceitos e aplicações dos métodos de ML na mineração e usar dados de um banco público para treino e teste de um algoritmo supervisionado de ML na descoberta de possíveis fatores envolvidos nas perdas e recuperação mássica de Ferro. O Ferro foi escolhido devido a sua importância para a mineração brasileira, especialmente no estado de Minas Gerais, e devido a disponibilidade de bancos de dados minerais contendo este elemento. Foi selecionado o algoritmo Random Forest (RF) e as análises foram realizadas através de códigos implementados em linguagem Python. O modelo treinado obteve acurácia geral de 74% na previsão de variáveis associadas à recuperação alta ou baixa de Ferro a partir de 13 variáveis de predição. Essa abordagem se mostrou uma forma rápida, sem custo e eficiente que pode fornecer várias informações importantes na elaboração de hipóteses relacionadas à mineração do Ferro. Ajustes no modelo podem conferir a maior acurácia esperada do RF e, o ajuste dos parâmetros de alta e baixa recuperação mássica de acordo com demandas de campo ou da indústria permitem a aplicação e ajustes do modelo aqui gerado em múltiplos contextos de pesquisa e industriais.

ABSTRACT

Mining is undergoing important changes made possible by the exponential technological development of the last decades. New environmental challenges and regulatory changes encourage the search for tools that enable more efficient and sustainable mining. In this context, the evolution of statistical and computational methods of artificial intelligence (AI) and machine learning (ML) are important tools that allow the massive analysis of high-resolution and large-scale data, allowing process improvement and increased productivity. An underexplored application is the use of AI to reduce losses during the beneficiation process. In iron mining, for example, it is estimated that losses in Brazilian plants can reach 20% and in Brazil 20 to 40% of the weight of the total ore mined is destined for tailings dams. Several factors are related to a lower-than-expected mass recovery, from variables associated with processing at the plant to those associated with the mineralogical composition of the site, however, establishing the contribution of each of these variables remains a challenge. Thus, the objective of this work is to review the literature on the concepts and applications of ML methods in mining and use data from a public bank to train and test a supervised ML algorithm in the discovery of possible factors involved in mass losses and recovery of iron. Iron was chosen due to its importance for Brazilian mining, especially in the state of Minas Gerais, and due to the availability of mineral databases containing this element. The Random Forest (RF) algorithm was selected and the analyzes were performed using codes implemented in Python language. The trained model had an overall accuracy of 74% in predicting variables associated with high or low iron recovery from 13 prediction variables. This approach proved to be a fast, cost-effective, and efficient which can provide several important information in the elaboration of hypotheses related to iron mining. Adjustments to the model can provide the highest expected accuracy of the RF, and the adjustment of high and low mass recovery parameters according to field or industry demands allow the application and adjustments of the model generated here in multiple research and industrial contexts.

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

A mineração e exploração de depósitos minerais está passando por mudanças importantes possibilitadas pelo exponencial desenvolvimento tecnológico das últimas décadas. Além disso, novos desafios ambientais e mudanças regulatórias fomentam a busca de ferramentas que possibilitem uma mineração mais eficaz e mais sustentável (FOSTER; GOLIYA, 2022; ODELL; BEBBINGTON; FREY, 2018).

Nesse contexto, a evolução dos métodos estatísticos e computacionais de Inteligência Artificial (IA) e aprendizado de máquina (*Machine Learning* - ML) como o aprendizado profundo de máquina (*Deep Learning* - DL) e Redes Neurais Artificiais (RNA) são ferramentas importantes na geoquímica e mineração atuais, pois permitem a análise de dados geofísicos de alta resolução e em grande escala, podendo gerar descobertas, aprimoramento de processos e ganhos de produtividade de forma complementar aos métodos de geoestatística tradicionais (SCHNITZLER; ROSS; GLOAGUEN, 2019; VERONESI; SCHILLACI, 2019). A abordagem através destas ferramentas tem ganhado cada vez mais espaço na mineração, permitindo análises exploratórias de baixo custo e em larga escala para ajudar a encontrar e avaliar características de depósitos minerais (HILL *et al.*, 2021).

Nas últimas duas décadas houve grande sucesso no uso de modelagem de dados e ML em algumas das áreas mais desafiadoras da exploração geoquímica, como elucidação de distribuições geoquímicas, prospecção mineral, predição de diferentes tipos de depósitos e a otimização de processos desde o beneficiamento até o aproveitamento de rejeitos (DEO *et al.*, 2021; ZHANG *et al.*, 2021; ZHAO; CHEN, 2021; ZUO, 2017, 2020; ZUO; CARRANZA, 2011; ZUO; ZUO, 2021).

Uma aplicação importante e ainda pouco explorada na mineração é o uso das ferramentas de IA e ML na redução de perdas da quantidade de minerais de interesse durante o processo de beneficiamento. Na mineração de Ferro, por exemplo, estima-se que as perdas nas usinas brasileiras geradas na fase de deslamagem do Ferro variem entre 8% até valores superiores a 20% (MATIOLO *et al.*, 2020). Vários fatores estão relacionados a uma recuperação mássica de Ferro inferior ao esperado, desde variáveis associadas ao beneficiamento na usina até fatores associados à composição

mineralógica do sítio na mina; no entanto estabelecer a contribuição de cada uma dessas variáveis permanece um desafio (GOMES; TOMI; ASSIS, 2015).

O aumento na produtividade é essencial para melhorar a sustentabilidade das operações (GOMES; TOMI; ASSIS, 2015; SAROUFIM, 2016). Neste contexto, é importante desenvolver processos que possam maximizar a recuperação de recursos. Até o momento, grandes esforços têm sido concentrados em melhorar o conhecimento das reservas de minério de Ferro e beneficiamento de Ferro (DEO *et al*., 2021; MERDITH *et al*., 2019; SHENG *et al*., 2015; TOHRY *et al*., 2022; ZHANG *et al*., 2021) e o uso de ML na otimização da recuperação mássica é uma ferramenta importante já usada para alguns minerais (FLORES; KEITH; LEIVA, 2020), mas ainda não explorada nessa aplicação para o Ferro. Essa abordagem pode se somar as técnicas de IA para obter melhor rendimento e sustentabilidade na mineração de minério de Ferro (SAROUFIM, 2016).

O minério de Ferro se tornou um dos principais produtos exportados pelo Brasil, que é o segundo maior exportador de minério de Ferro no mundo. Segundo os dados divulgados pelas Estatísticas do Comércio Exterior Brasileiro (COMEX STAT), em 2020, o mineral teve uma participação de 11,6% nas exportações totais e de 51,6% na indústria extrativa (NASCIMENTO, 2021).

Em 2017, de acordo com o Instituto Brasileiro de Mineração (IBRAM), o Brasil possuía a terceira maior reserva de Ferro do mundo, equivalente a 13,5% do total mundial (IBRAM, 2018). As exportações de minério de Ferro apresentaram uma variação anual máxima de 101,8% no período de 2000 a 2020, com uma média de US$ 12 bilhões nas vendas ao mercado externo (Figura 1) (BRASIL, 2020).

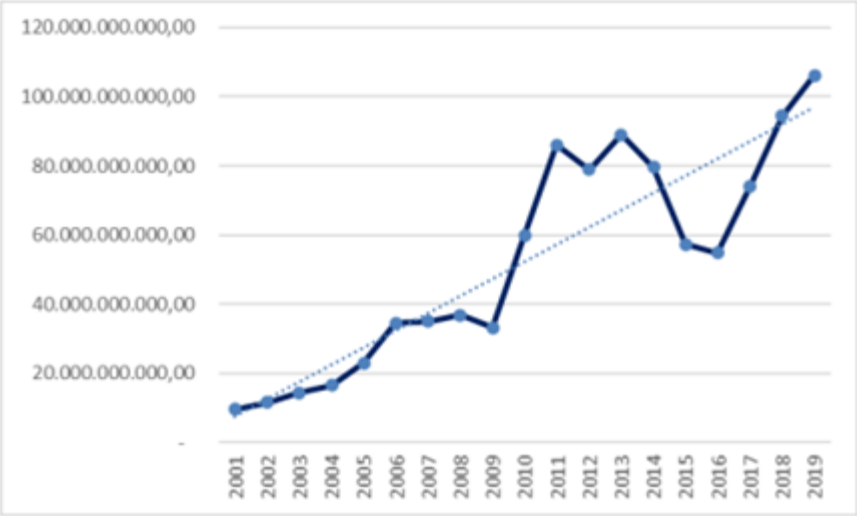


Figura 1: Produção de minério de Ferro no Brasil em bilhões de reais. **Fonte:** NASCIMENTO (2021).

Minas Gerais e Pará concentram quase 90% da produção brasileira de minério de Ferro (Figura 2), com destaque para a região do Quadrilátero Ferrífero (BRASIL, 2020). A mineração do Ferro envolve essencialmente dois tipos de resíduos sólidos: os estéreis e os rejeitos. Os estéreis são os materiais escavados que não possuem valor econômico e os rejeitos são resíduos resultantes dos processos de beneficiamento. A geração de rejeitos tem aumentado nos últimos anos, e o minério de Ferro é a substância que mais contribuiu para geração de rejeitos no Estado de Minas Gerais (IPEA, 2012).

O minério de Ferro brasileiro é classificado em dois tipos principais com base em sua mineralogia: a hematita e o itabirito. O minério do tipo hematita possui percentual médio de óxido de Ferro maior que 60% e é extraído principalmente no estado do Pará. Já o minério do tipo itabirito possui teores médios de 50% de óxido de Ferro e é extraído principalmente do estado de Minas Gerais, na região do Quadrilátero Ferrífero (IBRAM, 2015; IBRAM 2018). A escassez de depósitos com grandes concentrações de minério de Ferro faz com que os empreendimentos mineiros recorram a depósitos com teores mais baixos (SANTOS, 2018).

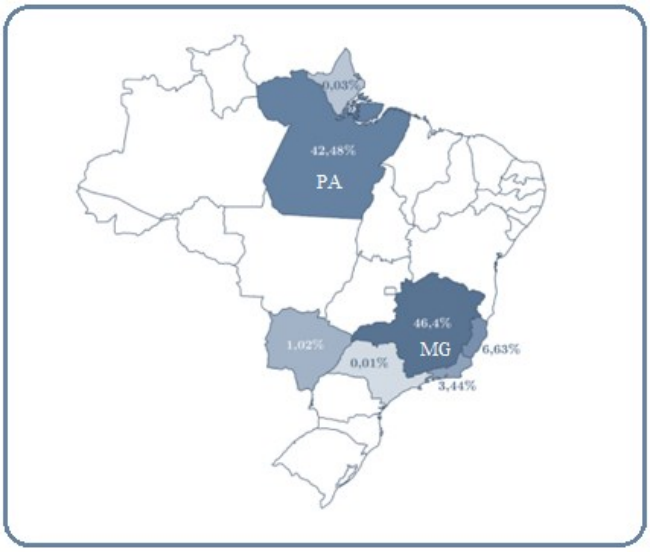


Figura 2: Principais estados brasileiros exportadores de minério de Ferro. **Fonte:** NASCIMENTO (2021).

Os minérios obtidos de ambas as reservas brasileiras são compostos predominantemente por hematita (αFe_2O_3), mas também possuem goethita (α-FeOOH) e magnetita (Fe_3O_4) em menores concentrações. Outros minerais, que não possuem Ferro, mas que ocorrem nestas formações minerais brasileiras são o quartzo (SiO_2) principalmente, a caulinita ($Si_2Al_2O_5(OH)_4$) e a gibbsita ($Al(OH)_3$) que introduzem o óxido de alumínio (Al_2O_3). Outros contaminantes químicos menores, como cálcio (Ca), magnésio (Mg), manganês (Mn), enxofre (S) e fósforo (P) também podem ser encontrados nos minérios brasileiros (YANG *et al.*, 2014).

Está amplamente documentado que óxido de Ferro, Sílica e Alumina são os principais constituintes dos rejeitos de minério de Ferro brasileiros, mas a composição e quantidade de Ferro nestes rejeitos varia enormemente (CARMIGNANO *et al.*, 2021), tanto entre os rejeitos brasileiros, expressivamente documentados a partir de rejeitos de mineradoras de Minas Gerais, quanto em relação a outros lugares do mundo (Tabela 1).

Tabela 1 - Composição química obtida por XRF (fluorescência de raios X) de diferentes rejeitos de minério de Ferro

Composição química principal %						
Fe_2O_3	**SiO_2**	**Al_2O_3**	**CaO**	**MgO**	**Outros**	**Localização da Mina**
8,38	90,40	0,43	0,06	< 0,1	0,63	Minas Gerais
11,6	84,20	1,60	-	-	2,6	Minas Gerais
11,31	75,23	2,64	1,47	2,10	7,25	Liaoning, China
12,31	34,72	16,22	7,63	8,92	20,2	Nanjing, China
15,1	84,4	0,45	0,07	< 0,1	0	Minas Gerais
18,58	36,48	11,67	16,85	5,66	10,76	Jiangsu, China
21,2	45,6	12,1	1,79	-	19,31	China
21,4	65,7	0,8	-	-	12,1	Minas Gerais
21,5	71,4	-	-	-	7,1	Minas Gerais
29,35	49,20	1,46	0,12	-	19,87	Minas Gerais
32,0	46,68	3,89	-	-	17,43	Minas Gerais
35,0	63,0	1,20	-	-	0,8	Minas Gerais
38,8	14	2,01	37,5	0,36	44,83	China
42,4	47,9	5,61	0,13	< 0,1	3,86	Minas Gerais
44,52	24,40	10,95	6,20	0,99	12,94	Hubei, China
47,80	30,0	21,2	0,1	0,1	0,8	Minas Gerais
51,37	15,11	3,39	0,23	0,16	29,74	Minas Gerais
55,78	16,58	15,46	1,44	0,13	10,61	Joda-Badbil, Índia
69,21	11,42	2,38	0,49	0,11	16,39	Bósnia Herzegovina
71,70	20,10	2,30	0,10	-	5,8	Minas Gerais
73,3	8,76	1,49	3,88	0,94	11,63	China

Fonte: modificado de CARMIGNANO et al. (2021)

O processo de beneficiamento de minério de Ferro, mostrado na Figura 3, envolve operações para modificar o tamanho das partículas e para aumentar o teor de Ferro, sem modificar as características químicas dos minerais (LIMA; ABREU, 2020). Após extraído, o minério de Ferro é lavrado, britado, moído, peneirado e separado de uma fase rica em

sílica, através do processo de flotação reversa (NAKHAEI; IRANNAJAD, 2017). Quando a concentração de óxido de Ferro é elevada, como no minério de Carajás, extraído no Pará, são realizadas apenas a britagem e separação por tamanho das partículas para se obter um produto pronto para comercialização (CARMIGNANO, 2021).

Já o minério do tipo Itabirito exige a adição de uma etapa de concentração – cujos principais métodos empregados são a separação magnética e gravitacional – que é realizada usando grande volume de água (CARMIGNANO, 2021). O uso de grandes volumes de água no processo de beneficiamento do minério de Ferro gera aumento expressivo no volume de rejeitos e, como consequência, multiplicam-se em quantidade e tamanho as barragens de rejeitos (LIMA; ABREU, 2020; ROY; NAYAK; RATH, 2020).

Ao final são obtidos os rejeitos grossos, gerados na fase de flotação, constituído principalmente de quartzo, e um rejeito fino, constituído principalmente de quartzo, hematita, goethita e caulinita. A fração fina do rejeito é gerada durante o processo de deslamagem, que acontece imediatamente após a flotação (SANTOS, 2018). No Brasil, é estimada uma geração de rejeitos (grossos e finos) entre 20% e 40% do peso do total do minério lavrado (LIMA; ABREU, 2020; MATIOLO *et al*., 2020).

Figura 3: Fluxograma típico do beneficiamento de minério de Ferro. Etapas de acordo com as Normas Brasileiras de Mineração. **Fonte:** CARMIGNANO (2021).

O Brasil é o segundo país no mundo que mais gera rejeitos na mineração de Ferro, atrás apenas da Austrália (Figura 4). No Brasil, entre 20% e 40% do peso do total do minério lavrado é destinado para barragens de rejeito (CARMIGNANO, 2021) e aproximadamente 95% dos rejeitos gerados na mineração de Ferro são destinados a 672 barragens (IBRAM, 2019), 340 das quais estão localizadas em Minas Gerais (FONTES *et al.*, 2019).

Entre 2015 e 2019, aconteceram mais de dez acidentes relacionados ao rompimento de barragens em vários países, consistindo em um problema grave, com enorme impacto econômico, ambiental e social (FALCONI *et al.*, 2020). A gravidade desse problema ficou evidenciada nos acidentes recentes de rompimento de grandes barragens em Minas Gerais: a barragem de Fundão, na cidade de Mariana, em 2015, que liberou mais de trinta milhões de metros cúbicos de água e sedimentos do tratamento de minério de Ferro, causando extenso dano ambiental e a morte de 19 pessoas (D'AZEREDO ORLANDO *et al.*, 2020) e o rompimento da barragem Córrego do Feijão, na cidade de Brumadinho em 2019 (THOMPSON *et al.*, 2020).

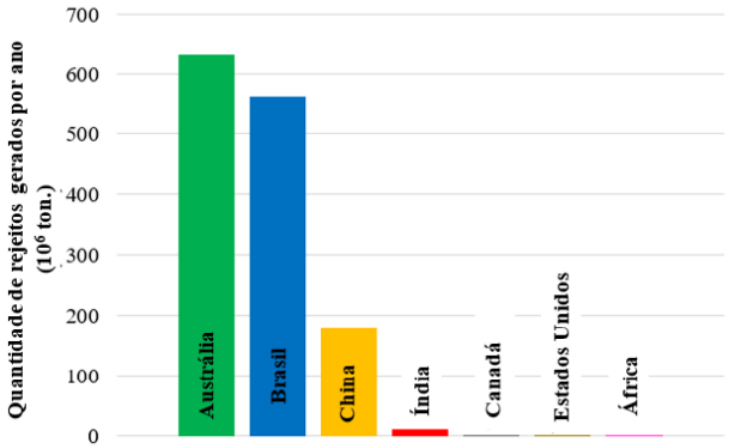


Figura 4: Quantidade de rejeitos de minério de Ferro gerado anualmente pelos principais países produtores. **Fonte:** CARMIGNANO (2021).

O aproveitamento dos rejeitos de minério de Ferro possui relevância local, pois outros países e os fornecedores globais ainda não têm como principal alvo tecnológico rejeitos de minério de Ferro (CARMIGNANO, 2021). Nesse sentido, há grande potencial

para a aplicação de tecnologias de ML associadas a mineração inteligente no Brasil. O uso de ML e tecnologias de IA pode melhorar o rendimento de processos, como a otimização da recuperação mássica do Ferro tanto durante o beneficiamento, como em materiais de rejeito da mineração, auxiliando na redução de contaminantes de rejeitos de minério de Ferro que geram grande impacto ambiental.

No entanto, as políticas públicas de inovação no setor de mineração no Brasil são percebidas como efêmeras, reativas e pouco estruturadas (BATISTA *et al.*, 2019) e os esforços, que se concentram em universidades e institutos de pesquisas, foram grandemente prejudicados nos últimos anos pela falta de investimento público e baixa colaboração do setor privado (CARMIGNANO, 2021).

Um levantamento em bancos de patentes com tecnologias para a transformação de rejeitos da mineração revelou que as patentes depositadas no INPI (Instituto Nacional de Propriedade Industrial do Brasil) priorizam ganhos nos processos e redução de custos e, apesar de algumas patentes se concentrarem em reaproveitamento de rejeitos, não foram observadas nestes processos o uso de nenhuma tecnologia de ML ou IA na base de inovação tecnológica avaliada (CARMIGNANO, 2021).

A necessidade de inovação no setor de mineração de Ferro e outros minerais é crescente em vista de desafios de sustentabilidade, aumento na demanda de exportação de Ferro para produção de aço em todo o mundo e redução das reservas. Dessa forma, nos últimos anos a indústria extrativa elevou seus gastos com Pesquisa e Desenvolvimento, o que sinaliza uma preocupação a respeito de como podem ser aprimorados os processos de extração e de produção, e há uma constante busca por aperfeiçoamento de técnicas para a indústria como um todo (CARMIGNANO, 2021; IBRAM, 2015; TOHRY *et al.*, 2022). Nesse sentido, a implementação de técnicas de ML na otimização de processos e aumento de recuperação mássica fornece um caminho promissor e ainda pouco explorado na mineração de Ferro no Brasil e pode contribuir para maior produtividade associada a mais sustentabilidade na cadeia de produção de minério de Ferro e outros minerais.

1.1 OBJETIVOS

Os objetivos deste trabalho são: i) fazer uma revisão da literatura referente aos conceitos e aplicações dos métodos de ML na mineração, sobretudo em relação à exploração de Ferro, e ii) aplicar o algoritmo de aprendizagem supervisionada *Random Forest* num estudo de caso utilizando dados de um banco de dados público para treinamento e subsequente análise de conjuntos de dados abertos para otimização de processos de mineração, visando aplicação no processo de recuperação mássica. Espera-se, com esta abordagem, contribuir com a integração interdisciplinar entre a ciência de dados, especificamente os métodos de ML, na mineração e geoquímica.

2. METODOLOGIA

Os conceitos, aplicações e exemplos levantados da literatura científica seguiram o modelo de revisão de literatura narrativa, portanto esta seção será dedicada a apresentar a metodologia empregada na análise de dados.

Em projetos de aprendizado de máquina é uma convenção usar um grande volume de dados cujo modelo originário é conhecido para treinar o modelo que irá fazer a análise de forma automatizada e avaliar a sua resposta (RODRIGUEZ-GALIANO *et al*., 2015). Uma fração destes dados é utilizada como um conjunto de dados de treinamento em si, para gerar modelos de predição ou regras associadas a um determinado conjunto de dados. O restante dos dados, denominado conjunto de teste, é usado para avaliar os parâmetros de previsão, a acurácia e a sensibilidade do modelo gerado (CHAOVALIT; ZHOU, 2005; CHAPELLE; SCHOLKOPF; ZIEN, 2009). O conjunto teste é apresentado ao modelo de ML gerado pelo tipo de algoritmo selecionado apenas após a conclusão da fase de treinamento. A precisão na comparação entre os resultados do modelo gerado com o conjunto de treinamento sobre o conjunto teste é considerada uma métrica precisa de como um modelo se comportaria em um novo conjunto de dados com características semelhantes (CATÉ *et al.*, 2017).

Para criar modelos do tipo classificação de dados é empregado o aprendizado de máquina supervisionado, amplamente usado para resolver desafios da mineração (CARRANZA; LABORTE, 2015; HILL *et al*., 2021; SCHNITZLER; ROSS; GLOAGUEN, 2019; SHENG *et al*., 2015). O aprendizado supervisionado reúne técnicas de ML aplicadas quando os dados estão no formato de variáveis de entrada (*input*) e valores de destino de saída (*output*) a partir dos quais o algoritmo aprende a função de mapeamento da entrada para a saída (ALLOGHANI *et al.*, 2020; CHAOVALIT; ZHOU, 2005).

Num modelo supervisionado, quando a variável de saída é uma de algumas categorias conhecidas, como animais, objetos e afins, chamamos esses algoritmos de classificadores. Quando a variável de saída é um valor real ou contínuo, como concentração de um mineral, são algoritmos de regressão (CHAPELLE; SCHOLKOPF; ZIEN, 2009; KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

Há ainda modelos mistos que podem combinar regressão e classificação, como o *Random Forest*, um modelo de aprendizado supervisionado do tipo *Ensemble*, termo geral dado para quando são usados múltiplos modelos em uma mesma tarefa ML e a resposta agregada de todos esses modelos é que será dada como o resultado final para cada dado que se está testando (PARMAR; KATARIYA; PATEL, 2019). Uma vez que o modelo é gerado, ele é aplicado ao conjunto de informações de entrada que se deseja testar (e.g., variáveis da composição mineralógica e geoquímica ao longo da mina, etapas do beneficiamento, parâmetros industriais) para prever informações de saída (e.g., região mais interessante para exploração, recuperação mássica esperada após o processamento, etapa do processo de beneficiamento que mais afeta perdas e outros parâmetros) (CARRANZA; LABORTE, 2015; HILL et al., 2021; SCHNITZLER; ROSS; GLOAGUEN, 2019; SHENG *et al*., 2015; VERONESI; SCHILLACI, 2019; ZHANG et al., 2021).

Para etapa de treinamento e análise de dados foi selecionada uma base de dados contendo parâmetros geoquímicos e de beneficiamento de Ferro. Para seleção do banco de dados foi escolhido um *dataset* pequeno (tanto em número de variáveis como número de amostras) para que a análise não dependesse do uso de servidores em nuvem e pudesse ser realizada num Notebook com processador Intel Core i7 1065G7, 12gb de memória RAM, GPU Intel Iris Plus integrada e Windows 10 Home 64bits, dispensando a necessidade de servidores ou computador de alto desempenho para ciência de dados. Foi selecionado um *dataset* que contivesse além de Ferro os outros minerais descritos como predominantes em composição similar aos minérios encontrados nos depósitos brasileiros (YANG *et al.*, 2014). Além disso, buscou-se um *dataset* que indicasse as concentrações dos diferentes minerais antes do beneficiamento e também após algumas das principais etapas de beneficiamento da mineração de Ferro no Brasil. O *dataset* escolhido foi o mineral ores around the world.csv, disponível para *download* (OLIVEIRA, 2022) e que contém as concentrações, convertidas em porcentagem, de Cr_2O_3, FeO, SiO_2, MgO, $Al2O_3$, CaO e P antes do beneficiamento e após os processos de deslamagem (*desliming mass recovery*), redução de alumina (*desliming Al reduction*) e perda por calcinação (LOI).

O algoritmo *Random Forest* foi usado para treinamento do *dataset* com objetivo de descobrir possíveis fatores envolvidos nas perdas e na recuperação mássica de Ferro após as etapas que equivalem à deslamagem e flotação.

O Ferro foi escolhido para o exercício demonstrativo neste trabalho devido à sua importância para a mineração brasileira e disponibilidade em bancos de dados públicos de *datasets* contendo dados das diversas etapas de prospecção e beneficiamento do mineral. Para fins de análise focada na mineração foram selecionados dados coletados experimentalmente em terra e a partir de furos de sondagem (CARRANZA; LABORTE, 2015; MERDITH *et al*., 2019).

O banco de dados foi tratado e analisado através de códigos implementados em linguagem Python 3.10.5 onde foram utilizadas as bibliotecas Pandas 1.4.3., Numpy 1.11, Scikit-learn 1.1 e Matplotlib 3.1 em conjunto com a aplicação Dataiku DDS (versão 11.0) (DATAIKU DSS 11.0), uma aplicação de inteligência artificial que suporta métodos escritos nas linguagens R e Python e é integrada por comandos *low-code* por meio de nódulos de informação, estruturados como um modelo *snowflake*. Foi seguido um fluxograma de pré-processamento dos dados (Figura 5), iniciando com a conversão de variáveis categóricas e/ou ordinais em variáveis numéricas, seguida da correção de valores nulos ou ausentes pelo método de substituição pelo valor da média (SCHÖNIG *et al.*, 2021).

A análise exploratória dos dados foi realizada conforme descrito por Sahoo et al. (2019) A biblioteca Pandas foi usada para importação e manipulação do *dataset*, a codificação *one-hot* foi usada para converter as variáveis categóricas em variáveis numéricas e os dados foram convertidos em *arrays* de entrada/saída. Em seguida, foram identificados quais parâmetros dentro da base de dados serviriam para identificar os alvos ou rótulos de saída (i.e., valores que se espera prever) e os recursos ou entradas (i.e., as colunas que o modelo usa para fazer uma previsão). Os *dataframes* gerados com o Pandas foram convertidos em *arrays* com a biblioteca Numpy e a base de dados foi dividida em um conjunto de treino e um conjunto menor de teste.

Durante o treinamento apresenta-se ao modelo tanto o conjunto de informações de entrada (concentração dos elementos convertida em %), quanto o conjunto de informações que se espera estar presentes na saída a partir dos dados de entrada, para

que ele possa aprender como criar uma regra que indique a alta ou baixa concentração de Ferro (medida de recuperação mássica). Para isso foi selecionado o algoritmo de aprendizagem supervisionada *Random Forest* importado da biblioteca Scikit-learn. Os dados de entrada foram divididos em 70% para treinamento e 30% para teste (STONE, 1974; XU; GOODACRE, 2018). Após finalizada a análise do conjunto de dados de treino, o algoritmo elabora o modelo ou o conjunto de regras que foi aplicado aos dados de teste e para avaliar a acurácia do modelo através de métricas de performance (LEE; ULLAH; WANG, 2020; MACHADO; MENDOZA; CORBELLINI, 2015). O fluxograma de trabalho está indicado na Figura 5.

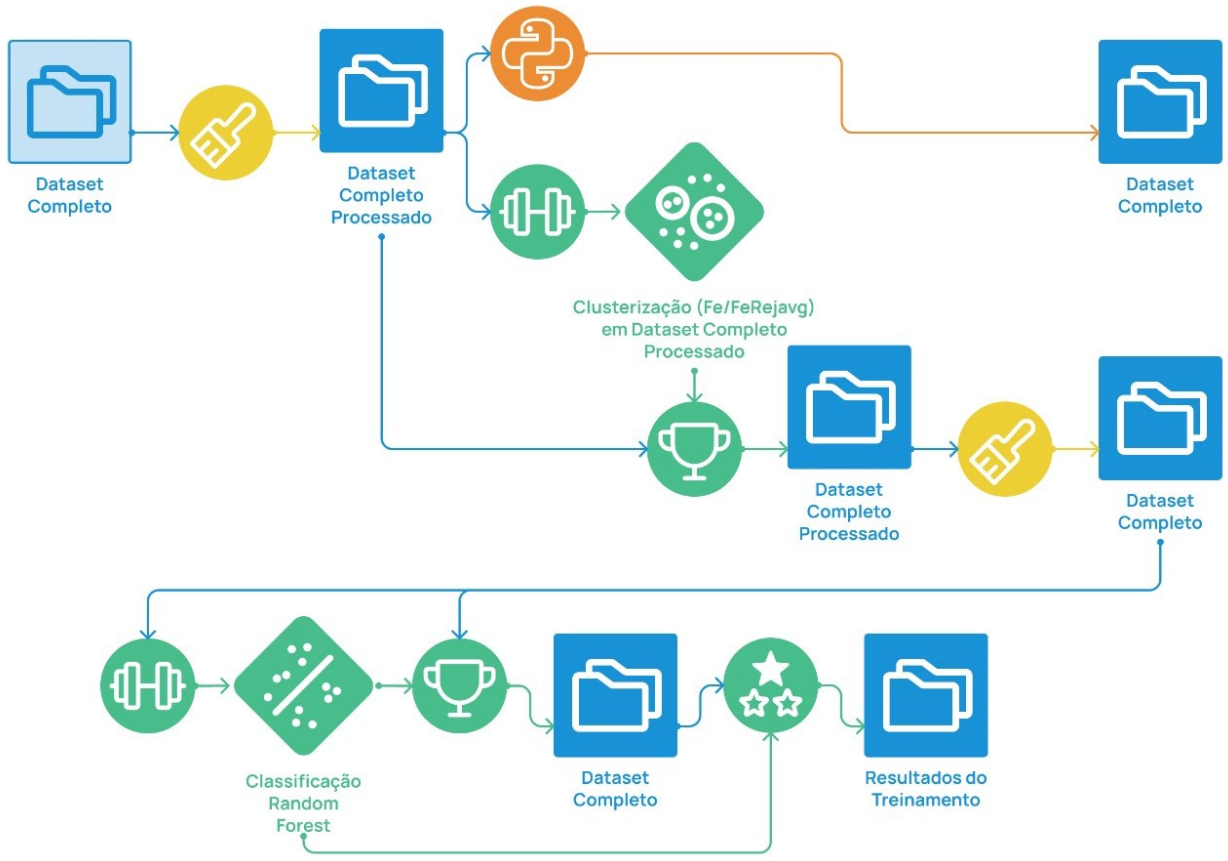


Figura 5: Fluxograma de trabalho realizado no Dataiku DSS.

Para as visualizações e geração de gráficos foi usada a biblioteca Matplotlib 3.5.2 (HUNTER, 2007). O código completo utilizado no trabalho está disponível no Anexo I.

2.1 Análise exploratória dos dados (AED)

A análise exploratória de dados é usada para ajudar a definir o comportamento das distribuições dos dados. Ela permite a manipulação de *datasets* complexos a partir de técnicas de manipulação de dados e análise estatística que descrevem a relação entre variáveis e hipóteses (SAHOO *et al.*, 2019).

Para isso são realizadas etapas como importação de um conjunto de dados, preparação e compreensão das variáveis e estudo das relações entre variáveis. A Figura 6 indica como a elaboração de hipóteses (*brainstorming*) está conectada com a compreensão das variáveis e como esta, por sua vez, está conectada novamente com a fase de *brainstorming*. Este processo gera *feedbacks* onde cada resposta obtida a partir de um dado pode funcionar para fazer novas perguntas (GOOD, 1983).

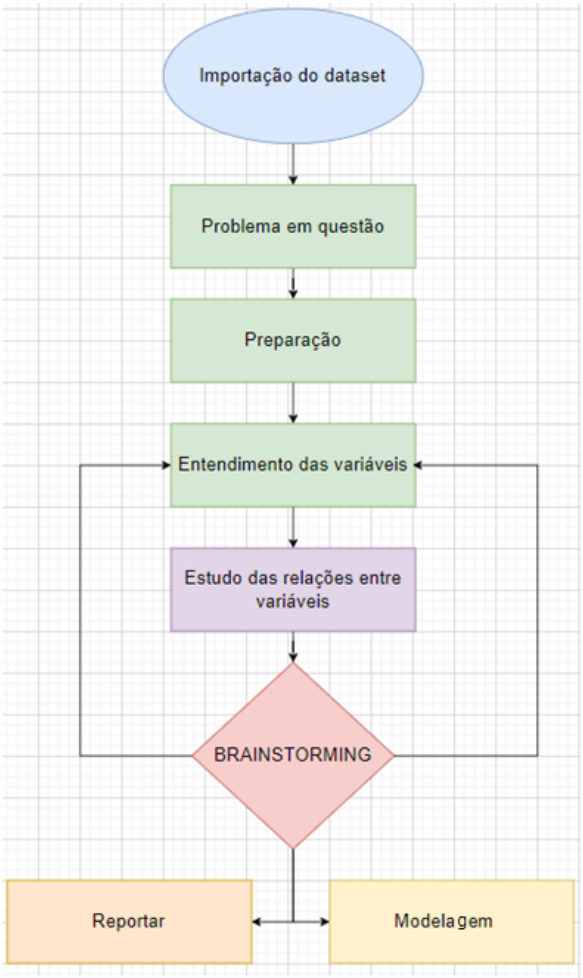


Figura 6: Fluxograma de análise exploratória de dados.

A AED foi realizada em linguagem Python. A abordagem geral adotada para a análise das distribuições das variáveis, bem como identificar os relacionamentos e correlações entre os teores de outros minerais do *dataset* descritos como co-ocorrentes em minérios de Ferro entre si e, em especial, com a variável teor de Ferro. Como forma de visualizar todas as relações entre as variáveis simultaneamente foi gerada uma matriz de correlação, que permite estabelecer uma relação quantitativa de dependência entre as diferentes variáveis utilizadas na análise.

A fase de EDA começa imediatamente após o conjunto de dados ser importado no início do fluxo de análise de dados. A importação de um conjunto de dados foi realizada com o Pandas, por meio de funções dedicadas à leitura dos dados (REBACK *et al.*, 2022). O arquivo contendo o conjunto de dados utilizado é o arquivo geologia_litotype.csv, um *dataset* com conjunto de variáveis lito-geoquímicas (teores em porcentagem de Cr_2O_3, FeO, SiO_2, MgO, $Al2O_3$, CaO e P) (OLIVEIRA, 2022).

2.2 Preparação dos dados

Nesta fase é realizada a limpeza do *dataset* para eliminar variáveis que sejam desnecessárias para o escopo da análise proposta (foram eliminados os elementos considerados traço em amostras de mineração de Ferro), remover dados redundantes, remover colunas duplicadas, ajustar a nomenclatura, adicionar novas variáveis etc. Os valores nulos foram substituídos pelo valor da média para a variável em questão (ZHANG *et al.*, 2006).

Após a preparação do *dataset* foi iniciada a exploração propriamente dos dados via análise bivariada e multivariada. A análise combinada da distribuição dos dados serve para revelar a presença de regiões de interseção entre duas ou mais variáveis. Correlações nas distribuições dos dados de concentração dos outros minerais em relação ao Ferro podem sugerir que, ao menos dentro de um intervalo de concentração, a presença daquele mineral se relaciona a maior ou menor concentração de Ferro recuperado, sugerindo uma possível variável explicativa.

Para entender a distribuição das variáveis numéricas foi realizado o comando **describe()**, a função retorna o resumo estatístico do *dataframe* ou série que inclui a

quantidade de variáveis, média, mediana (ou 50º percentil), desvio padrão, valores mínimos e máximo percentil das colunas. Quando a função de descrição é aplicada a um objeto de série, o resultado também é retornado na forma de série.

2.3 Visualização e Estudo da Relação entre as Variáveis

Uma matriz de dispersão bivariada foi elaborada para buscar possíveis correlações entre diferentes elementos químicos. As informações são distribuídas em uma grade de eixos de forma que cada variável numérica nos dados seja compartilhada nos eixos y em uma única linha e nos eixos x em uma única coluna (Figura 28). Este tipo de visualização é útil para observar os relacionamentos mais importantes de forma simultânea. No entanto, essa é uma função computacionalmente árdua, por isso é mais adequada para conjuntos de dados com um número relativamente baixo de variáveis (BEHRENS, 1997; SAHOO *et al.*, 2019). A matriz de dispersão foi gerada a partir de funções da biblioteca Seaborn conforme mostrado no código do Anexo I.

A AED também permite a validação das análises da matriz de dispersão através da análise de distribuição de cada conjunto de dados através de histogramas. Apesar da distribuição normal não ser um requisito para o algoritmo Random Forest (RF) (PARMAR; KATARIYA; PATEL, 2019), essa análise é particularmente relevante quando se busca fazer inferências sobre os dados através de métodos paramétricos.

A normalidade e distribuição dos dados também é eficientemente avaliada através da dispersão de probabilidades da melhor linear, em que se ajustam a tendência dos dados obtidos através de regressão que represente a reta que melhor ajuste a tendência dos dados. O uso deste recurso permite inferir sobre a confiança dos dados presentes no *dataset* em relação aos modelos probabilísticos, pois quanto melhor o ajuste, maior a confiança dos dados em modelos paramétricos (seguem uma distribuição normal), (JORDAN; ZHANG; HIGGINS, 2007) como as correlações simples e análises de covariância.

A matriz de dispersão é um tipo de análise de covariância amplamente empregada durante a AED, pois ela permite a visualização rápida de correlações entre as múltiplas variáveis do dataset, permitindo identificar semelhanças e relações positivas ou

negativas entre duas variáveis, assim como a ausência de correlação. O coeficiente de correlação é limitado à linearidade e, portanto, não se aplica a nenhuma relação não linear. Na AED estes dados podem ser indicados tanto pelas distribuições dos pontos (amostras) das variáveis, que é interessante para observar a relação entre variáveis onde há intervalos que se correlacionam (como dentro de uma faixa de concentração de dois elementos) ou como valores numa matriz de covariância que apenas indica relação negativa, positiva ou ausência de correlação.

2.4 Treinamento e teste do modelo empregando o *Random Forest*

Após a análise exploratória, foi realizado o treinamento do modelo com divisão 70-30% do *dataset* e realizada a validação cruzada usando *v-fold* simplificado (STONE, 1974; XU; GOODACRE, 2018). O código completo desta etapa encontra-se no Anexo I.

Para o treinamento foram criadas as classes lógicas **True** (valor numérico 1) a ser atribuída se o valor da concentração de Ferro no rejeito (*ferej_map*) associada ao parâmetro testado (*desliming mass recovery*, *desliming Al reduction*, Fe, Si, P, Al, Mn, Ti, *Loss on Ignition*/LOI, Ca, Mg, K e Na) fosse considerada alta ou **False** (valor numérico 0) caso fosse baixa. O valor de referência utilizado para classificar a concentração em uma dessas classes lógicas foi o valor médio do parâmetro *ferej_map* (média da variável Ferro) para então gerar a matriz de confusão, que é uma maneira de expressar quantas das previsões feitas por um classificador estavam corretas.

A matriz de confusão é uma medida usada na resolução de problemas de classificação, aqui aplicada para classificação multiclasse, representando a contagem de valores previstos e reais. Na matriz de confusão são apresentados pares de relações entre variáveis, em que valores próximos a 0.0 indicam que a variável preditora não está relacionada à alta recuperação mássica e valores próximos a 1.0 indicam variáveis preditoras relacionadas com altos índices de recuperação mássica.

Após o treinamento feito a partir de 70% dos dados presentes no *dataset*, obtém-se a regra de classificação gerada pelo modelo, a qual é então aplicada ao conjunto teste equivalente aos 30% do *dataset* que não foram utilizados no treinamento. A partir disso foram obtidas as métricas de performance deste modelo, indicadas por índices

numéricos para quantificar estas métricas. Para esta análise foram considerados os seguintes parâmetros:

i) Precisão: indica o índice de acertos para cada classe 1 e 0 (alta e baixa recuperação mássica de Ferro, respectivamente);

ii) Suporte: informa quantas variáveis foram agrupadas em cada classe 1 e 0;

iii) *Recall*: corresponde à sensibilidade ou taxa de verdadeiros positivos e indica a fração de todos os valores **True** reais que foram classificados corretamente, o que corresponde à fração de parâmetros que verdadeiramente estão relacionados à recuperação mássica e que foram classificadas corretamente pelo modelo;

iv) Acurácia: indica a porcentagem de previsões que foram feitas corretamente e é obtida pela razão do número de previsões que o modelo acertou em relação ao número total de previsões. Este parâmetro é um bom indicador do equilíbrio entre as classes, visto que valores de Suporte próximos entre si correspondem a um bom equilíbrio entre as classes;

v) F1-Score: é a média harmônica de *recall* e precisão, e em um conjunto de dados desequilibrado capturará melhor o desempenho do modelo em cada classe.

3. REVISÃO DE LITERATURA

A Inteligência Artificial é o campo da computação que usa algoritmos computacionais para criar máquinas com inteligência inspirada na inteligência humana (ANGELOV *et al.*, 2021; FLEMMER *et al.*, 2009). Dentre os vários métodos de IA há aqueles que fazem o computador simular um processo de aprendizado por si só a partir de dados de treinamento e que são chamados de *Machine Learning* (ML) ou aprendizado de máquina (MICHALSKI; CARBONELL; MITCHELL, 1983). Neste método de IA o computador aprende por observação de exemplos de forma similar à que o cérebro humano aprende (GOSWAMI, 2021).

A técnica de ML que se inspira no aprendizado do cérebro humano, uma das mais poderosas e amplamente usadas técnicas de ML, é chamada de aprendizado profundo ou *Deep Learning* (DL) (LECUN; BENGIO; HINTON, 2015). DL é baseado na forma como os neurônios biológicos se comunicam e a base do DL são as Redes Neurais Artificiais (RNA) (FRANCOIS-LAVET *et al.*, 2018). Nessa técnica são empregados algoritmos com arquitetura, ou seja, camadas de processamento, que mimetizam o funcionamento dos neurônios humanos (NIELSEN, 2015).

A definição original de ML é: “um programa de computador aprende com a experiência E correspondente a uma classe de tarefas T e medida de desempenho D, se seu desempenho nas tarefas T, medido por D, melhora com a experiência E” (MICHALSKI; CARBONELL; MITCHELL, 1983).

A Figura 7 indica a relação hierárquica entre IA, ML, DL e RNA desenvolvidas e estudadas no campo da ciência denominado Ciência de Dados (CARPENTER et al., 2018).

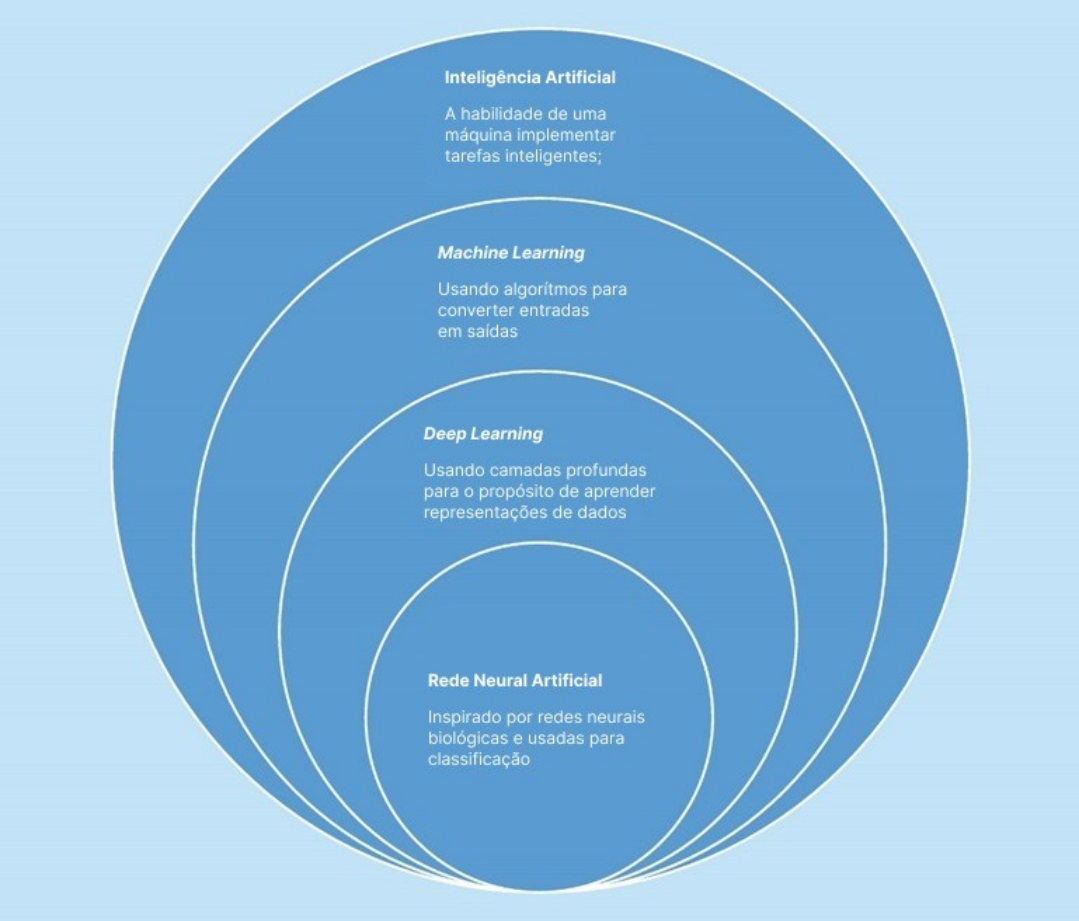


Figura 7: Representação da hierarquia entre diferentes conceitos relacionados ao aprendizado artificial. Esses conceitos existem como subconjuntos uns dos outros: inteligência artificial (IA), *Machine Learning* (ML), *Deep Learning* (DL) e Redes Neurais Artificiais (RNA). **Fonte:** modificado de CARPENTER et al. (2018).

A ideia de criar uma arquitetura computacional que mimetizasse o funcionamento dos neurônios surgiu nos anos 40, inspirada em como essas células especializadas se comunicam e processam informações para resolver problemas computacionais complexos (BRAGA; CARVALHO; LUDERMIR, 2007).

Os neurônios biológicos são compostos por dendritos que recebem uma informação, o núcleo que processa a informação e o axônio que passa a informação para outro neurônio (GREENGARD, 2001). Na década de 1950, Rosenblatt desenvolveu o conceito matemático de uma rede neural que consistia em uma única camada de neurônios McCulloch-Pitts, aos quais ele se referiu como Perceptrons (ROSENBLATT,

1958). Na década de 60, foi proposta uma implementação computacional para o neurônio, o Modelo Perceptron, que mimetiza os neurônios biológicos (HAYKIN, 2001) de forma matemática através de operadores condicionais lógicos e retorna saídas binárias (0 ou 1, equivalentes a Falso e Verdadeiro). Na sua forma mais simples (Figura 8) o Perceptron possui três camadas: uma camada de entrada que recebe a informação de forma semelhante à que um neurônio biológico recebe a informação pelos dendritos, uma camada de processamento que processa a informação tal como o núcleo do neurônio, e uma camada de saída que transmite a informação de forma análoga ao que ocorre no axônio do neurônio biológico. Como essas redes possuíam uma única camada de processamento, o seu aprendizado era bastante limitado. Embora se soubesse que essa limitação poderia ser superada usando mais de uma camada, a forma de ajustar os pesos das redes conectados à(s) camada(s) oculta(s) só foi resolvido na década de 1970, com o algoritmo de retropropagação (*backpropagation*) derivado por Werbos em 1974 e redescoberto por Rumelhart et al. (1986) (RUMELHART; HINTON; WILLIAMS, 1986).

Atualmente, vários outros algoritmos também são usados para o treinamento de RNA com múltiplas camadas (NIELSEN, 2015; SCHMIDHUBER, 2015). Assim como o cérebro tem alto poder de processamento devido à combinação de múltiplos neurônios biológicos, os neurônios artificiais também são unidos em uma espécie de rede, daí o nome rede neural, que conecta um neurônio artificial a outro (HAYKIN, 2001).

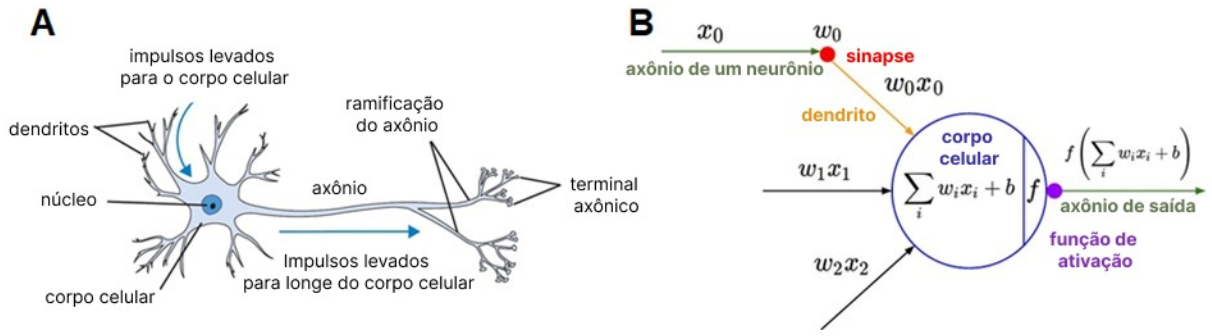


Figura 8: Comparação entre um neurônio biológico (A) e um neurônio ou nó de uma RNA (B). **Fonte:** modificado de CS231N, [s.d.].

Para uma rede neural funcionar ela precisa de dados de entrada e dados de saída (NIELSEN, 2015). A descrição matemática que relaciona a(s) saída(s) do sistema com sua(s) entrada(s) usa uma equação diferencial calculada a partir de experimentos no sistema e pode ser desenvolvida sem uma compreensão detalhada da ciência do sistema (MILITKÝ, 2011). Após a especificação da estrutura de um modelo experimental, os dados de entrada e saída do processo são utilizados para calcular os parâmetros da equação diferencial usando um método de identificação paramétrica (DOHERTY, 1999).

A função de um modelo é generalizar informações sobre um determinado sistema (MILITKÝ, 2011). Em um sistema hipotético F_0 causal haverá um conjunto de causas (x_0) informadas na entrada e um dado efeito observado na saída (y_0). É possível que distúrbios ocorram inserindo erros (E_0) no processo e gerando variáveis aleatórias na saída y_0 (Figura 9), ou conforme Militký define:

> "A modelagem é uma maneira de descrever algumas das características de um sistema investigado (original) usando um modelo (físico, abstrato) e critérios definidos. Em vez das entradas x0, um subconjunto de variáveis explicativas x é usado e as saídas y0 são substituídas pela resposta escalar y. As funções desconhecidas F0 que transformam entradas em saídas são substituídas pelo modelo f(x, β). As perturbações E0 são caracterizadas por erros εi (erros devidos à medição). A seleção da melhor forma de modelo é o principal objetivo da modelagem" (MILITKÝ, 2011, p. 46).

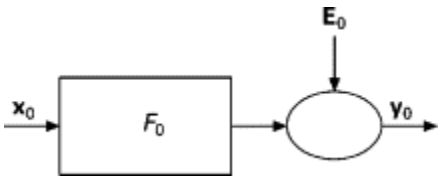


Figura 9: Ilustração de um sistema determinístico com perturbações estocásticas. **Fonte:** (MILITKÝ, 2011).

De acordo com a área de estudo, os modelos podem ser simples e bem-organizados ou grandes e mal organizados, quando há muitas variáveis e nem todas elas são conhecidas ou suas influências não podem ser completamente medidas (MOLUGARAM *et al.*, 2017). Em disciplinas técnicas, como as geociências, normalmente encontra-se sistemas do tipo difusos e parcialmente desorganizados, em que há

processos físicos e mensuráveis envolvidos, mas há fatores e conexões desconhecidos ou parcialmente conhecidos que influenciam os sistemas (LARY, 2010).

Nestas disciplinas comumente são empregados modelos empíricos construídos com relação à capacidade de previsão ou ajuste do modelo (aproximação de dados), capacidade de prognóstico (previsão) e estrutura do modelo (concordância com teorias e fatos) (MILITKÝ, 2011). Os modelos empíricos determinísticos oferecem soluções simplistas para comparações quantitativas entre diferentes condições, variáveis e erros que influenciam um sistema e são obtidos a partir da análise de correlações de dados experimentais (BADEDA *et al.*, 2017).

Os modelos empíricos e determinísticos mais comumente usados nas ciências técnicas usam processos de ajuste de curvas para generalizar os resultados dos experimentos (COTTIS, 2012; LARY *et al*., 2016). Alternativamente, métodos de ML como as redes neurais têm sido usados para construir modelos empíricos aplicados as diversas áreas das geociências (ALAVI; GANDOMI; LARY, 2016; LARY *et al.,* 2016; MEREMBAYEV; YUNUSSOV; YEDILKHAN, 2019) ou ainda a criação de novos modelos determinísticos combinados com técnicas de ML. O uso de empirismos construídos a partir de aprendizado de máquina em um modelo baseado em dados físicos resulta em um modelo híbrido (GOLDSTEIN; COCO, 2015; MIN; YOON, 2021).

Em muitos casos, não é possível construir um modelo matemático com base nas informações disponíveis sobre o sistema sob investigação. Nesses casos, uma abordagem interativa para a construção de modelos pode ser atraente (GOLDSTEIN; COCO, 2015; MILITKÝ, 2011). Embora modelos empíricos determinísticos possam fornecer uma maneira eficiente de modelar dados complexos, Goldstein e Coco (2015) destacam que:

> “Qualquer modelo que exija expressões empíricas também está sujeito a revisão à medida que a parametrização empírica é refinada: todo empirismo está aberto à revisão por dados corroborantes ou conflitantes. À medida que mais dados se tornam disponíveis e mais graus de liberdade são explorados, torna-se mais difícil incorporar todos os dados disponíveis em um único preditor empírico ideal [...] parametrizações empíricas para modelos numéricos devem ser construídas usando técnicas de aprendizado de máquina porque essas técnicas são construídas para operar em grandes conjuntos de dados de alta dimensão” (GOLDSTEIN; COCO, 2015, p. 1).

Algoritmos de ML se tornaram difundidos nas geociências nas últimas décadas e sua aplicação na mineração é frequentemente associada aos avanços tecnológicos na indústria 4.0 (JUNG; CHOI, 2021). A incorporação de métodos e conceitos de IA pela indústria da mineração é chamada de mineração inteligente, que se desenvolve a partir da demanda de análise, interpretação e aplicação de informações geradas a partir de uma grande quantidade de dados continuamente produzida, coletados e compartilhados em tempo real (CHOI; LEE, 2020).

É previsto um investimento de 218 milhões de dólares pelas empresas de mineração em plataformas de IA em todo o mundo até 2024, com uma taxa de crescimento anual composta de 23,4%, com destaque para a maior penetração de tecnologias de IA das minas da Australásia e africanas (GLOBAL DATA, 2019).

O uso das tecnologias de IA e ML na mineração pode ser empregado tanto em novas análises em dados de prospecção e de pesquisa geoquímica regional já depositados em bancos de dados privados e públicos (BRITISH COLUMBIA RGS, 2020), até dados coletados em tempo real ou armazenados em sensores e objetos inteligentes (Internet das Coisas, ou IoT – *Internet of Things*) desde durante perfurações (QISHUAI *et al.*, 2018; ZARE *et al.*, 2019) até o processo de beneficiamento (SAROUFIM, 2016). Esse volume e complexidade de dados se beneficia das chamadas análises *big data*, onde algoritmos de ML são frequentemente empregados (Figura 10) (CHAKRABORTI, 2022).

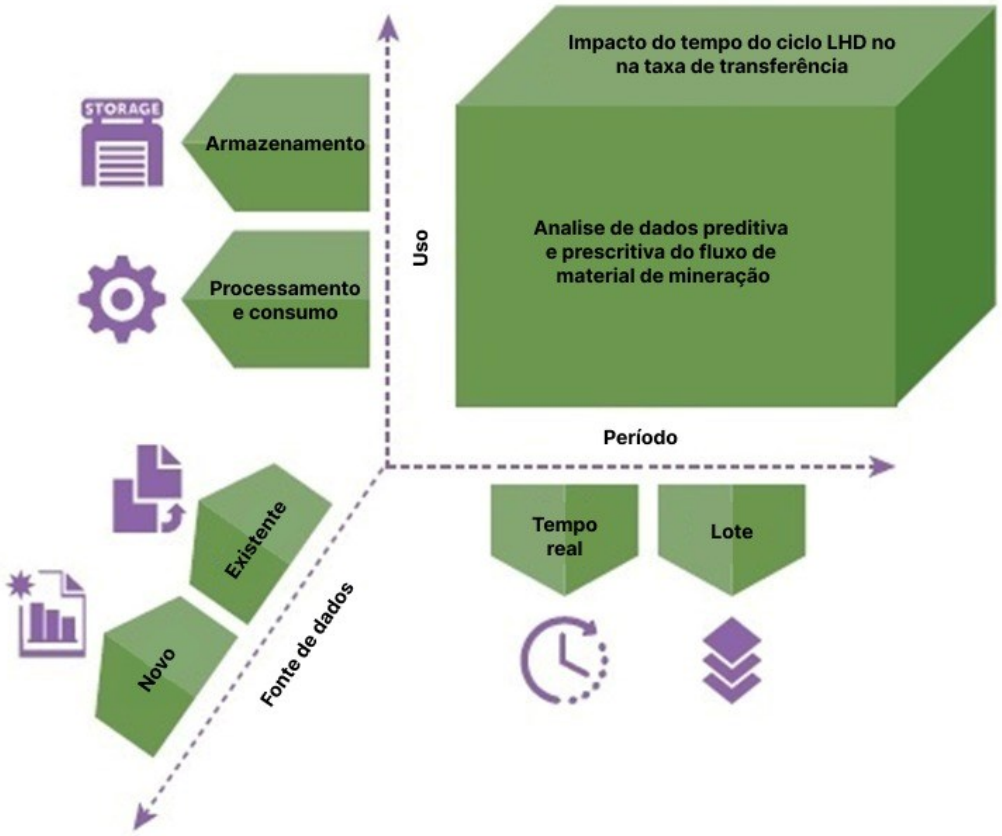


Figura 10: Aplicação de soluções de *big data* para análises de mineração inteligente. **Fonte:** modificado de CHAKRABORTI (2022).

Algoritmos de ML são aproximadores universais (HORNIK; STINCHCOMBE; WHITE, 1989). O teorema de aproximação universal afirma que uma rede neural com uma única camada oculta, contendo um número finito de neurônios, pode aproximar qualquer função contínua se algumas suposições sobre a função de ativação forem atendidas (CSÁJI, 2001).

Esse teorema é a base do aprendizado de máquina (ML) realizado pelas redes neurais artificiais e explica a possibilidade mais amplamente usada dessa tecnologia: aprender o comportamento subjacente de um sistema a partir de um conjunto de dados de treinamento e aplicá-lo a conjuntos de dados mais amplos (CATÉ *et al.*, 2017; KRATSIOS; BILOKOPYTOV, 2020).

Outra característica interessante das técnicas baseadas em ML é que elas não precisam de um conhecimento prévio sobre a natureza das relações entre os dados e

podem extrair informações a partir destes de forma automatizada, permitindo que sejam usadas com diferentes fins e para resolver múltiplos tipos de questões dentro da geoquímica e da mineração (CATÉ et al., 2017; LARY, 2010; LARY *et al*., 2016). Lary (2010) classifica as aplicações de ML em geociências em três tipos:

(1) O modelo determinístico do sistema tem alta demanda computacional e ML pode ser usado como uma ferramenta aceleradora de código.
(2) Não existe um modelo determinístico, mas um modelo empírico baseado em ML pode ser criado a partir dados existentes.
(3) Problemas de classificação.

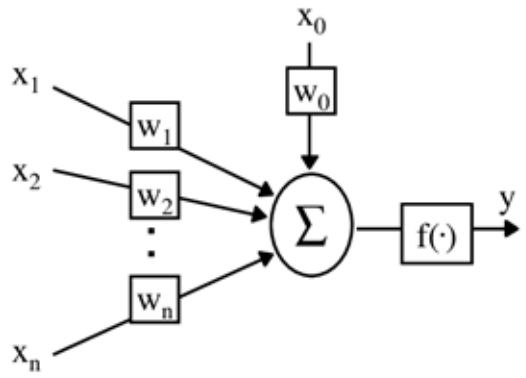
Esta seção apresenta uma descrição da arquitetura das Redes Neurais Artificiais e dos tipos de Aprendizado de Máquina, destacando-se em seguida algumas das aplicações de ML em mineração e geoquímica encontradas em trabalhos acadêmicos a partir das buscas em bancos de dados de artigos especializados e focada nas aplicações 2 e 3 citadas na relação acima, uma vez que elas mostraram-se de emprego recorrente nos trabalhos revisados e cobrem os dois principais tipos de uso de ML: o uso de algoritmos de ML devido à sua capacidade de regressão (aplicação 2), o que permite tanto prever o valor da variável dependente da saída quando alguma informação sobre as variáveis explicativas está disponível quanto estimar o efeito de alguma variável explicativa sobre a variável dependente, e a capacidade de classificação (aplicação 3) que permite obter rótulos com denominações específicas a partir das variáveis explicativas (LARY, 2010; LARY *et al*., 2016).

3.1 Arquitetura das Redes Neurais Artificiais (RNA):

Uma RNA é composta por várias unidades de processamento conectadas por estruturas de comunicações às quais se atribui um determinado peso, de forma que o aprendizado nesse tipo de arquitetura advém das interações entre as unidades de processamento da rede (NIELSEN, 2015).

O tipo de RNA mais simples e comumente usada é a *feed-forward*, em que as conexões entre os nós não formam um ciclo, i. e., uma camada se conecta à camada seguinte sem que a informação tenha um caminho de volta. A unidade básica de uma

RNA *feed-forward* é referida como um nó, neurônio ou elemento de processamento (SCHMIDHUBER, 2015) baseado no modelo de neurônios de McCullock-Pitts (1943). Este modelo (Figura 11) constitui uma soma ponderada e forma a base da arquitetura da maioria das redes neurais, embora uma função de ativação contínua atualmente seja mais comumente usada do que a função binária original (DOHERTY, 1999; SCHMIDHUBER, 2015).



$$y = f(\mathbf{x} \cdot \mathbf{w}) = f(\sum_{i=0}^{n} w_i x_i)$$

Figura 11: O neurônio de McCulloch-Pitts e sua equação equivalente. O neurônio consiste em uma soma ponderada linear de suas entradas calculadas por uma função de limite binário em que as entradas estão representadas por x, seus pesos respectivos por w e a saída por y. **Fonte:** DOHERTY (1999).

Numa RNA os sinais (X1, X2, Xn) são apresentados à entrada e cada sinal é multiplicado por um valor, ou peso (W1, W2, Wn), que indica a sua influência na saída da unidade e, em seguida, é realizada a soma ponderada dos sinais que produz um nível de atividade (Figura 11). Se este nível de atividade exceder um certo limiar (*threshold*) a unidade produz uma determinada resposta de saída (NIELSEN, 2015).

A arquitetura de rede é a com que os neurônios de uma RNA estão estruturados. Ela consiste em uma camada de entrada, uma ou mais camadas intermediárias (camadas ocultas) e uma camada de saída (HORNIK; STINCHCOMBE; WHITE, 1989; NIELSEN, 2015). Na primeira camada se faz a entrada dos dados apresentados como padrões à rede. São denominados preditores, ou seja, são os exemplos do que se quer que a máquina aprenda. O papel dos neurônios desta camada é distribuir as entradas a todos os neurônios da próxima. A camada de entrada também

é responsável por calcular qual o peso de cada conexão com a camada intermediária (predição) (HAYKIN, 2001). Em seguida estão as camadas ocultas, que podem ser em número de uma ou mais e são as responsáveis pela maior parte do processamento das informações. Os padrões enviados pela camada de entrada são decodificados, ou seja, têm suas características analisadas pelos neurônios da(s) camada(s) intermediária(s) e o resultado é passado para a camada de saída (NIELSEN, 2015). Os padrões obtidos nas camadas ocultas têm os resultados concluídos e apresentados na camada de saída. Ela recebe os estímulos da camada intermediária e constrói o padrão que será a resposta. Este processo é ilustrado no diagrama da Figura 12 e o fluxo da informação se faz da esquerda para a direita.

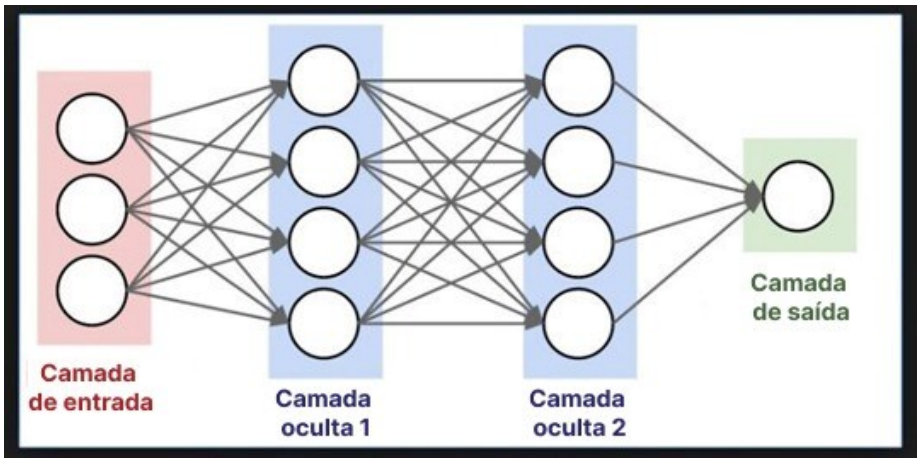


Figura 12: Uma rede neural de 3 camadas com três entradas, duas camadas ocultas consistindo em quatro neurônios cada e uma camada de saída. **Fonte:** modificado de CS231N, [s.d.].

O objetivo de um algoritmo de treinamento é permitir que a RNA represente um modelo ou regra que descreva o comportamento de entrada/saída de um sistema não linear. Para isso, o algoritmo tenta minimizar uma função de custo ajustando os parâmetros de peso da RNA (NIELSEN, 2015). A função de custo é usada no contexto de problemas de otimização para atribuir intuitivamente valores de uma ou mais variáveis num número real representando algum "custo" associado ao evento (BARNARD; WALD, 1953). Ela é uma medida de quão bem a RNA se ajusta a um conjunto de padrões de dados de treinamento de entrada/saída que o sistema produziu (SCHMIDHUBER, 2015),

correspondendo, portanto, a uma função de análise de desempenho do processo computacional.

A propriedade mais importante das redes neurais é a habilidade de aprender a partir de entradas fornecidas e com isso melhorar seu desempenho através do processo de ajustes aplicado a seus pesos, o que é chamado de treinamento (NIELSEN, 2015). O aprendizado ocorre de forma bem-sucedida quando a RNA consegue chegar a uma solução generalizada para uma classe de problemas e é enfim gerado um modelo de aprendizado, que é um conjunto de regras bem definidas para a solução de um problema de aprendizado (SCHMIDHUBER, 2015).

3.2 Tipos de Aprendizado de Máquina

A forma como os pesos nas camadas de uma RNA são modificados determina o tipo de algoritmo a ser utilizado no aprendizado. Essa abordagem, o tipo de dados que são usados como entrada e saída, e o tipo de problema que eles resolvem, irão organizar os algoritmos de aprendizado em quatro classes principais de ML: supervisionado, não supervisionado, aprendizado por reforço; métodos híbridos, como os algoritmos semi-supervisionados, *Ensemble* (comitê de classificação) e D*eep Learning* também são empregados (SAH, 2020).

3.2.1 Aprendizado supervisionado

Esta classe de aprendizado é aplicada quando os dados estão no formato de variáveis de entrada e valores de destino de saída e o algoritmo aprende a função de mapeamento da entrada para a saída, a qual é apresentada como grupos de dados rotulados (Figura 13) (SAH, 2020).

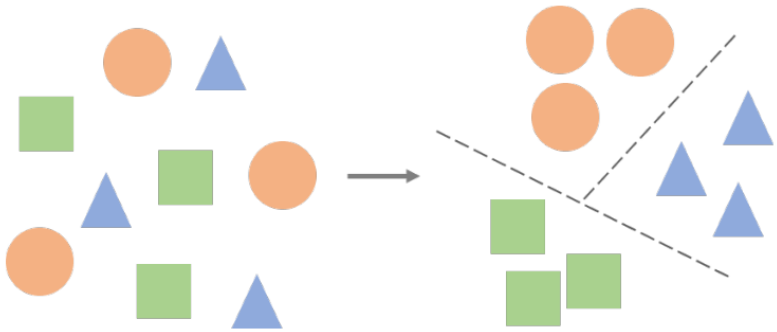


Figura 13: Visão geral da aprendizagem supervisionada. Exemplos de entrada são categorizados em conjuntos específicos conhecidos como classes. **Fonte:** modificado de SAH (2020).

Esses algoritmos são usados especialmente quando há grandes conjuntos de dados e podem ser divididos em algoritmos de classificação se a variável de saída y é contínua ("positivo" ou "negativo", "bola" ou "xis") ou algoritmos de regressão quando y é uma variável numérica contínua (Figura 14) (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

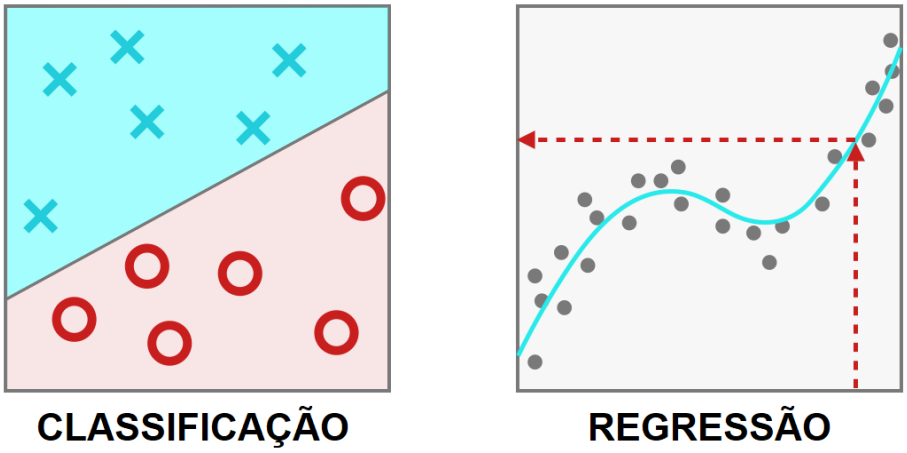


Figura 14: Tipos de aprendizado supervisionado. A linha separa os dados classificados como X ou O (resultado de um algoritmo de classificação) ou o ajuste da linha fornece uma previsão de saída para um dado de entrada fornecido (resultado de um algoritmo de regressão). **Fonte:** Modificado de sharpsightlabs.com (EBNER, 2021).

As máquinas de vetor de suporte (*Support Vector Machines* - SVM) são um tipo popular de algoritmo supervisionado usado principalmente em problemas de

classificação e de regressão (SCHÖLKOPF, 1998). Os algoritmos SVM são baseados na teoria da dimensão VC (Vapnik-Chervonenkis), uma medida da capacidade de um conjunto de funções que podem ser aprendidas por um algoritmo de classificação binária (CORTES; VAPNIK; SAITTA, 1995). São usados especialmente para resolver bases de dados pequenas e não lineares no reconhecimento de padrões de alta dimensão (CERVANTES *et al.*, 2020).

O objetivo dos SVMs é encontrar um hiperplano que otimize a aresta entre os dois pontos amostrais mais próximos (Figura 15). Um hiperplano é o limite de decisão que permite classificar quais amostras ficarão em cada grupo definido para a classificação dos dados. Os hiperplanos podem gerar limites para segregar as classes no espaço n-dimensional de várias formas diferentes, dependendo do modelo de classificação adotado. Os pontos de amostra no limite de borda maximizado são chamados de vetores de suporte e o meio seção da aresta é o hiperplano de classificação ideal, o ajuste da curva (NOBLE, 2006).

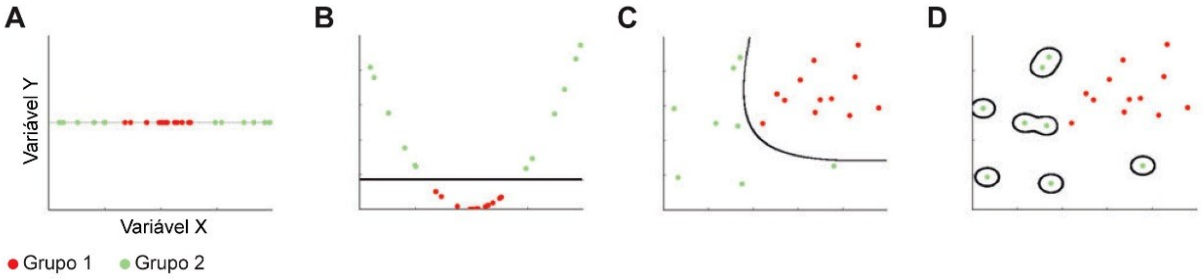


Figura 15: Diferentes formas de ajuste das curvas de SVMs. **a)** Um conjunto de dados unidimensional não separável. **b)** Hiperplano separando previamente dados bidimensionais inseparáveis. **c)** Um conjunto de dados bidimensional linearmente não separável, que é linearmente separável em quatro dimensões. **d)** Um SVM que superajustou um conjunto de dados bidimensional. **Fonte:** modificado de NOBLE (2006).

3.2.2 Aprendizado não supervisionado

Esta classe de aprendizado é empregada quando os dados de entrada x estão disponíveis na base de dados, mas as variáveis de saída y não estão disponíveis, i.e., os dados não estão devidamente rotulados. Desta forma, a base de dados é trabalhada

para o modelo aprenda mais sobre suas características (SAH, 2020). A principal classe de algoritmo não supervisionado é o agrupamento, ou *clustering*, em que, como o nome indica, os dados de entrada são classificados em estruturas (*clusters*) que podem ser mais facilmente compreendidos e manipulados de acordo com suas similaridades e divergências (Figura 16) (BENGIO *et al.*, 2009).

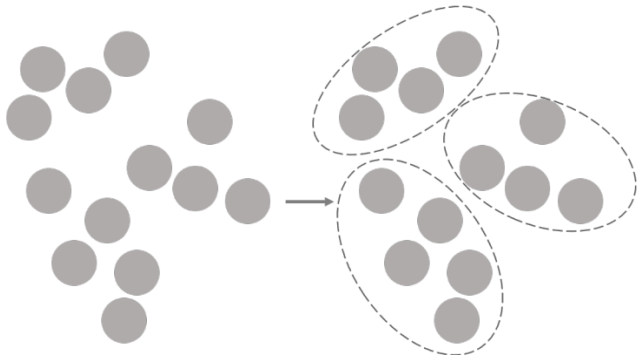


Figura 16: Visão geral do aprendizado não supervisionado. As amostras de entrada são agrupadas em *clusters* com base em padrões identificáveis. **Fonte:** modificado de SAH (2020).

3.2.3 Aprendizado por reforço

Classe de aprendizado empregada quando um agente externo avalia a resposta fornecida pela rede em tarefas de tomada de decisão, através de um sistema que recompensa as ações corretas e penaliza as ações incorretas (Figura 17), buscando maximizar a recompensa total (SAH, 2020) para que o aprendizado seja concluído com sucesso.

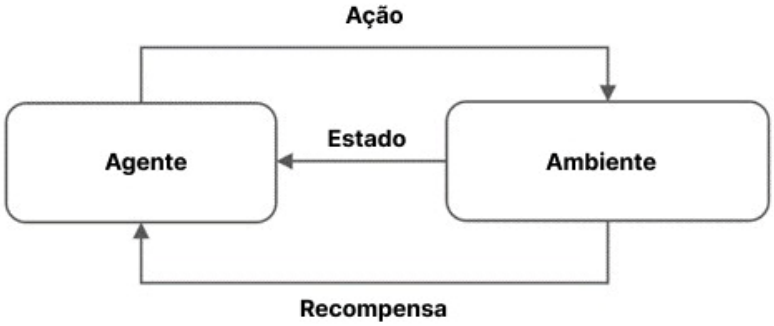


Figura 17: Visão geral do aprendizado por reforço. Um agente observa o estado do ambiente e executa ações para maximizar uma recompensa. **Fonte:** modificado de SAH (2020).

3.2.4 Métodos híbridos

Frequentemente os métodos não supervisionados são combinados com os supervisionados, uma abordagem híbrida denominada de aprendizado semi-supervisionado (Figura 18), em que os dados de treinamento contêm muito poucos exemplos rotulados e muitos exemplos não rotulados (SAH, 2020). Dados similares são agrupados inicialmente através de um algoritmo não supervisionado criando grupos rotulados que, em seguida, são usados para rotular o restante dos dados não rotulados existentes na base de dados (CHAPELLE; SCHOLKOPF; ZIEN EDS., 2009).

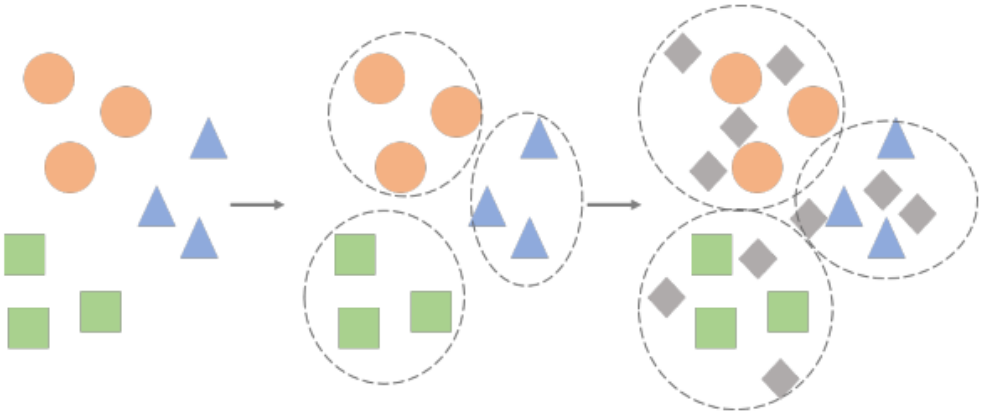


Figura 18: Visão geral do aprendizado semi-supervisionado. Os aglomerados formados por uma grande quantidade de dados não rotulados são usados para classificar uma quantidade limitada de dados rotulados. **Fonte:** modificado de SAH (2020).

Um método híbrido frequentemente aplicado nas geociências são os algoritmos de comitê de classificação (*Ensemble learning*) (HE *et al.*, 2022). Nesses algoritmos são usados métodos de agrupamento com vários algoritmos treinados em conjunto (*ensembles*) para resolver o mesmo problema e obter melhor desempenho preditivo do que seria obtido por qualquer um dos algoritmos de aprendizado caso estes fossem utilizados isoladamente (Figura 19) (POLIKAR, 2012). Esse método é particularmente importante para conjuntos de dados complexos e é um dos mais usados no sensoriamento remoto, geoquímica aplicada e mineração (HE *et al.*, 2022; JUNG; CHOI, 2021; LARY, 2010). Nesse método a capacidade de generalização de um comitê *ensemble* é geralmente muito mais forte do que o da base de aprendizes (SAH, 2020). O método *Random Forest*, utilizado no presente estudo, encaixa-se nesta categoria.

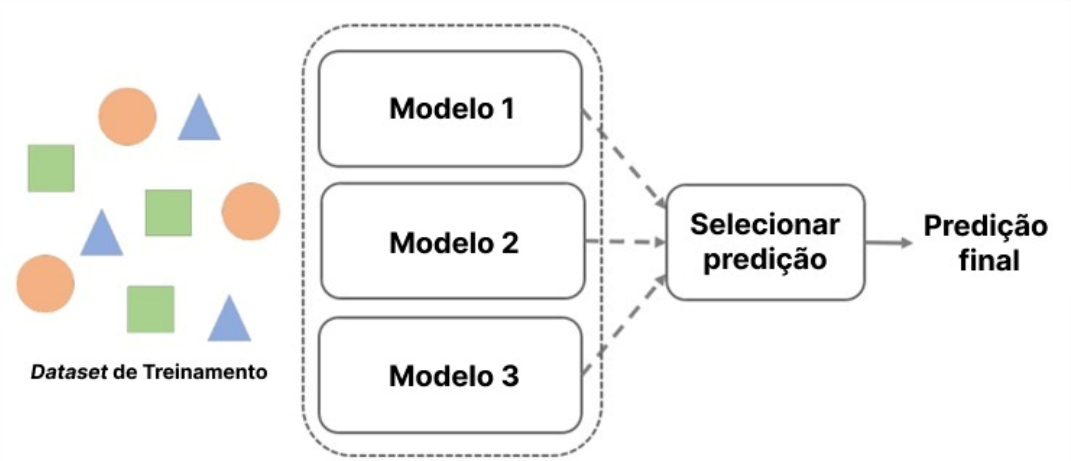


Figura 19: Visão geral da aprendizagem *Ensemble*. Vários modelos são treinados para a mesma tarefa e suas previsões individuais são usadas para obter o melhor resultado, com base num sistema de votação. **Fonte:** modificado de SAH (2020).

Por fim, as RNA já citadas anteriormente podem combinar métodos de aprendizado em múltiplas camadas, através da associação entre vários Perceptrons ou nós em redes de aprendizado profundo ou *Deep Learning* (DL) (FRANCOIS-LAVET *et al.*, 2018). O DL pode ser supervisionado, não supervisionado, semi-supervisionado, auto supervisionado ou de reforço, dependendo principalmente de como a rede neural é usada (LECUN; BENGIO; HINTON, 2015).

O aprendizado profundo é um subcampo do ML e as redes neurais compõem a espinha dorsal do aprendizado profundo e é o número de camadas/nós, ou profundidade, de redes neurais que distingue uma única RNA de um algoritmo de DP, que possui mais de três camadas (Figura 20) (AIZENBERG; AĬZENBERG; VANDEWALLE, 2000).

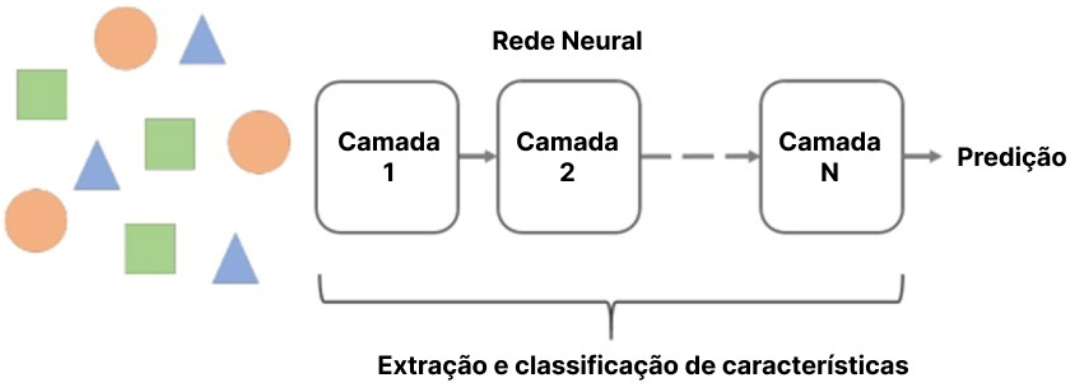


Figura 20: Visão geral do aprendizado profundo. Rede neural multicamada que aprende a extração e classificação de recursos de ponta a ponta. **Fonte:** modificado de SAH (2020).

As redes neurais de *Deep Learning* são amplamente empregadas no reconhecimento de padrões e as redes neurais convolucionais (RNC) são a classe mais comumente aplicada para analisar imagens (FRANCOIS-LAVET *et al.*, 2018). As RNC realizam o aprendizado repetindo a operação e enfatizando as características dos dados de entrada através de uma operação de convolução. Detalhes de como funcionam os algoritmos de DL foram revisadas por LeCun, Bengio e Hinton (2015), que explicam como esse método revolucionou vários campos do conhecimento e tecnologias:

> "O aprendizado profundo permite que modelos computacionais compostos de várias camadas de processamento aprendam representações de dados com vários níveis de abstração. [...] O aprendizado profundo descobre uma estrutura complexa em grandes conjuntos de dados usando o algoritmo de retropropagação para indicar como uma máquina deve alterar seus parâmetros internos que são usados para calcular a representação em cada camada a partir da representação na camada anterior. Redes convolucionais profundas trouxeram avanços no processamento de imagens, vídeo, fala e áudio, enquanto as redes recorrentes iluminaram dados sequenciais, como texto e fala" (LECUN; BENGIO; HINTON, 2015, p. 436).

Um exemplo de aplicação importante de *Deep Learning* nas operações de mineração é a automatização da identificação dos tipos de minerais presentes numa amostra antes do estágio de processamento mineral através de dados de imagens hiperespectrais, por exemplo (Figura 21). Como o padrão espectral é único para cada

mineral, ao combinar imagens hiperespectrais DL para identificar rapidamente os tipos de minerais contidos em rochas usando um método não destrutivo (OKADA *et al.*, 2020).

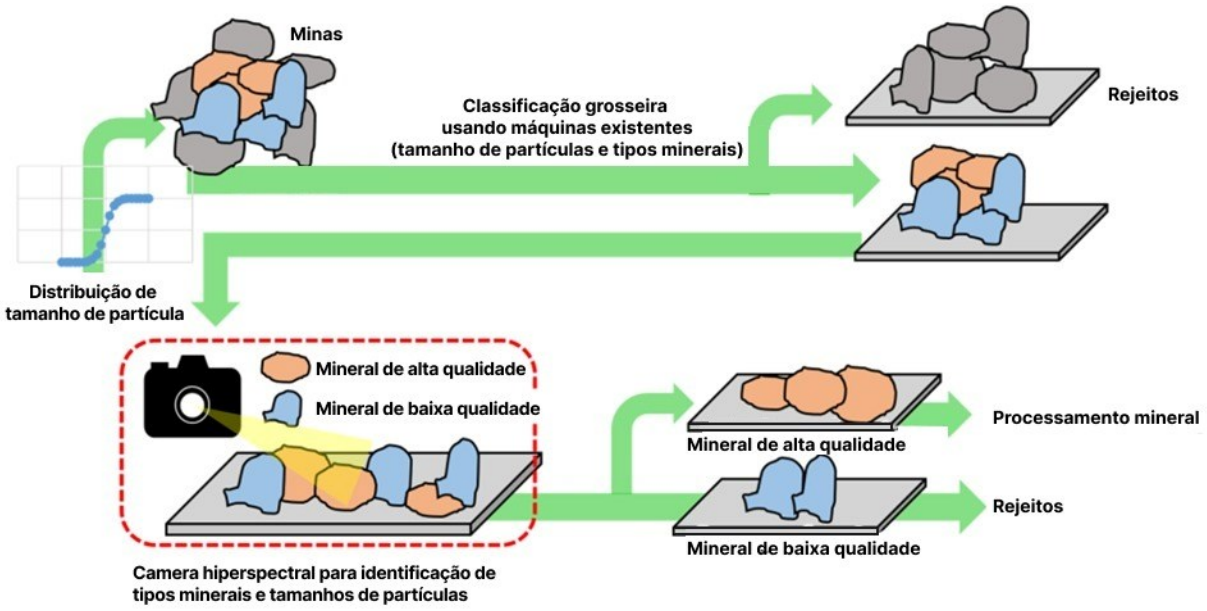


Figura 21: Diagrama esquemático do processamento mineral usando imagens hiperespectrais e *Deep Learning*. Após a lavra do minério e classificação granulométrica e de cor são coletadas imagens hiperespectrais e algoritmos de DL são empregados para identificar e classificar os minerais de alto grau de pureza para processamento e os rejeitos, o que possibilita operações mais eficientes. **Fonte:** modificado de OKADA et al. (2020).

Okada et al. (2020) usaram um método de aprendizado profundo baseado em aprendizado supervisionado e classificação numa RNC capaz de extrair automaticamente as características dos dados de entrada a partir de um banco de dados de imagens hiperespectrais (Figura 22). A partir do espectro mineral de entrada, a RNC automaticamente reconheceu, extraiu e aprendeu as formas espectrais que são consideradas específicas de cada mineral, permitindo uma classificação rápida e automatizada (SINAICE *et al.*, 2017). A análise de dados hiperespectrais usando DL permitiu a identificação do grupo de minerais estudado com acurácia de mais de 90% (OKADA *et al.*, 2020).

As RNCs possuem várias camadas de convolução, que calculam operações de convolução várias vezes. A convolução é uma operação linear que envolve a multiplicação de um conjunto de pesos com a entrada, bem como uma rede neural tradicional. Como a técnica foi projetada para entradas de imagens bidimensionais, a multiplicação é realizada entre um *array* de dados de entrada e um *array* bidimensional de pesos, chamado de filtro ou Kernel (CHEN; GUO; LI, 2020). No exemplo da figura abaixo, há entradas de imagens hiperespectrais de diferentes minerais. As imagens hiperespectrais contém informações de todo espectro eletromagnético, então as entradas equivalem aos valores de pixel de cada espectro, para cada mineral, nos diferentes comprimentos de onda. Nas camadas de convolução, todos os pixels em seu campo receptivo são convertidos em um único valor e normalizados (OKADA et al., 2020).

A saída final da camada convolucional é um vetor que é transferido e processado novamente em camadas de agrupamento (ou *pooling*) que simplificam a informação de saída da camada convolucional anterior (LI *et al.*, 2021), resultando num mapa de características condensadas (Figura 22).

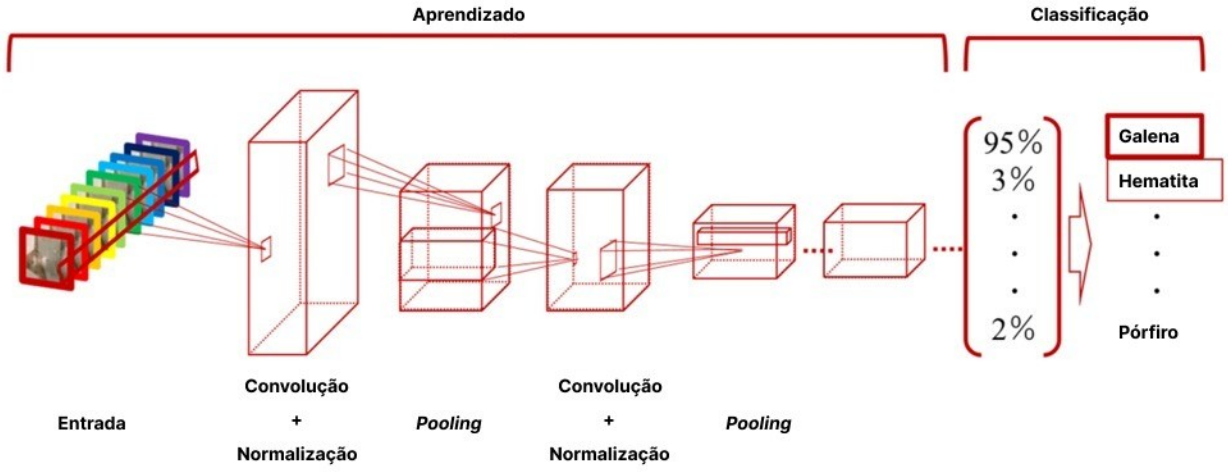


Figura 22: Estrutura da uma Rede Neural Convolucional (RNC) para dados hiperespectrais. Os dados de saída são agrupados com rótulos de resposta correta e incorreta. A rede neural consiste em dupla camada de convolução e normalização seguida de *pooling* usadas para aprender quais parâmetros hiperespectrais correspondem a cada mineral de entrada. Em seguida é realizada uma classificação para

identificação de cada mineral (saída) correspondente a cada entrada (perfil hiperespectral). **Fonte:** OKADA et al. (2020).

3.3 Aplicações dos principais tipos de ML na Mineração

Em uma ampla revisão sistemática recente, Jung e Choi (2021) analisaram 192 estudos em que métodos de ML foram aplicados na mineração e mostraram, de forma detalhada, como as pesquisas estão distribuídas, os tipos de base de dados que são usados nestas publicações e os tipos de algoritmos usados. Os autores elaboraram um mapa técnico do setor de mineração incluindo as subcategorias onde são usadas ferramentas de ML nas três fases do processo de mineração: prospecção/exploração, extração/operação (lavra e beneficiamento) e recuperação/restauração. Os modelos usados mais frequentemente foram SVM (58 estudos), *Deep Leaning* (56 estudos) e *Ensemble* (56 estudos) (JUNG; CHOI, 2021).

Após a classificação, 109 trabalhos que usaram ML na mineração foram divididos em 34 campos (Figura 23). Dentre os 59 trabalhos encontrados por Jung e Choi (2021) que usaram ML na etapa operacional de extração, as principais aplicações foram na detecção de anomalias geoquímicas usando análises exploratórias (ZHANG *et al.*, 2019), estudos de mapeamento geológico (BERETTA *et al.*, 2019; BÉRUBÉ *et al.*, 2018; COSTA; ORTALE; RITACCO, 2011), análise mineral (ACOSTA et al., 2019; HOOD; CRACKNELL; GAZLEY, 2018; RAHMAN et al., 2015, 2016; SCHNITZLER; ROSS; GLOAGUEN, 2019) e estudos de mapeamento e prospecção mineral usando dados exploratórios, aplicação que concentrou o maior número de trabalhos revisados (CARRANZA; LABORTE, 2015a, 2015b; GRANEK; HABER, 2015; LI; CHEN; XIANG, 2019; RODRIGUEZ-GALIANO *et al*., 2015; SUN *et al.,* 2019; TESSEMA, 2017; XIONG;

ZUO; CARRANZA, 2018; ZHANG; ZHOU; LI, 2018; ZUO; CARRANZA, 2011).

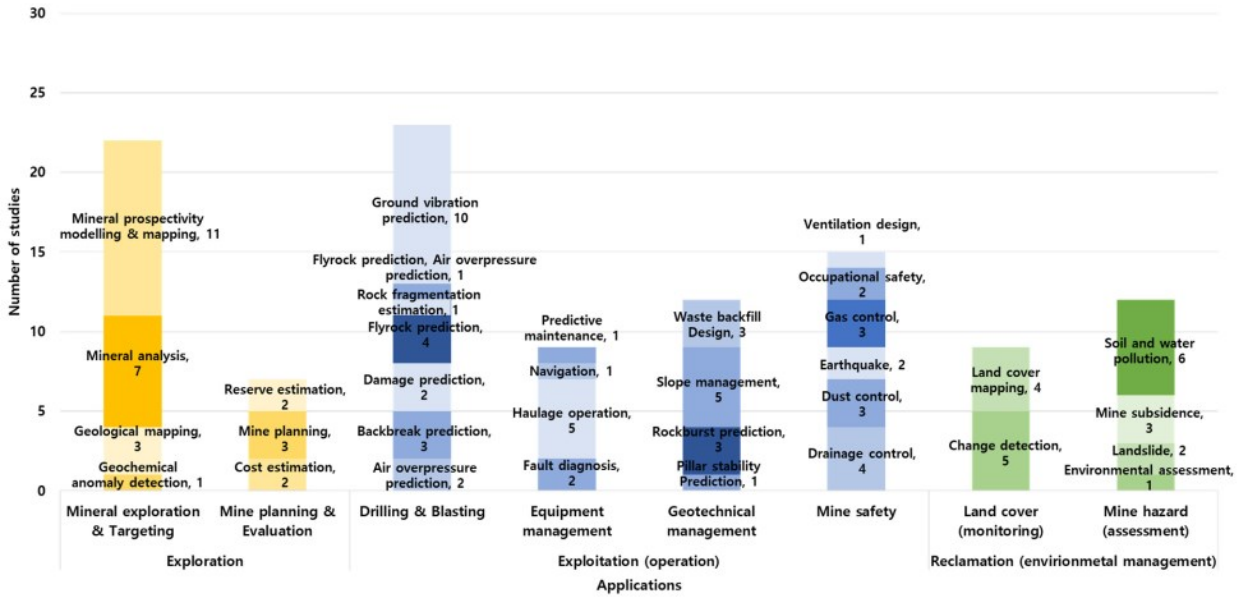


Figura 23: Campos de aplicação de ML nas diferentes fases do processo de mineração e número de publicações em cada categoria. **Fonte**: JUNG e CHOI (2021).

A maior parte dos artigos incluídos na revisão se concentraram na etapa operacional, assistindo diretamente nos processos de extração de minérios (Figura 24) (JUNG; CHOI, 2021).

Figura 24: Número de estudos usando ML em cada fase da mineração. **Fonte**: modificado de JUNG e CHOI (2021).

A revisão sistemática também apontou que a maior parte dos trabalhos utilizou bancos de dados abertos (bancos públicos como RapidEye, NASA, Minto, etc), seguido

de dados de campo e laboratoriais, respectivamente (Figura 25a). Dentre as formas de validar os algoritmos de ML para avaliar a acurácia do modelo, os mais usados foram bancos de dados de teste (Figura 25b) (JUNG; CHOI, 2021). Os dados de validação são usados para fins de verificação para selecionar o modelo mais adequado entre todos os modelos de ML. Os dados de teste são usados para determinar quão bem o modelo de ML selecionado funciona (WILLMOTT, 1982).

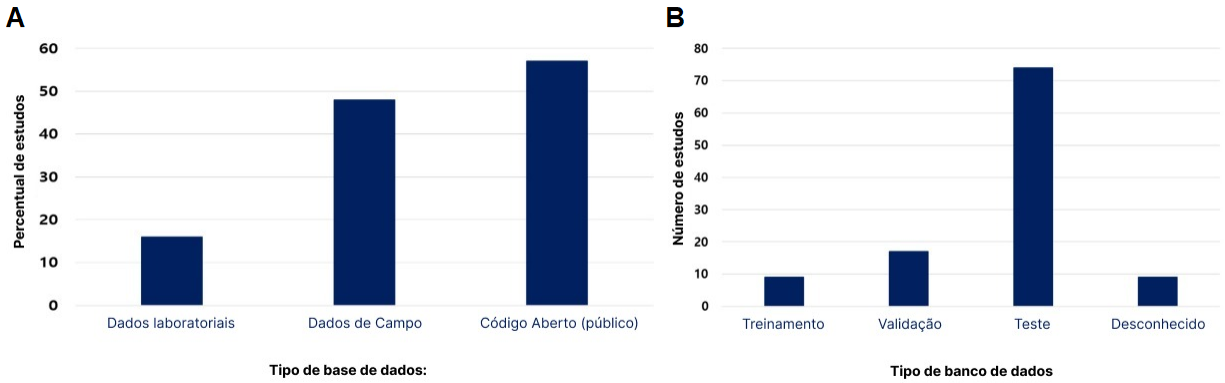


Figura 25: Fontes de informação mais usadas em aplicações de ML na mineração. **a)** Tipos de base de dados. **b)** Tipos de bancos de dados. **Fonte**: modificado de JUNG e CHOI (2021).

O trabalho de Jung e Choi (2021) também examinou as frequências de uso de cada modelo de ML em cada campo da mineração (Tabela 2). Na etapa de exploração mineral, os métodos de *Ensemble* e SVM foram os mais utilizados e empregados especialmente para prospecção e sondagem, além de também serem amplamente usados na etapa de extração. O DL foi usado principalmente nas etapas de perfuração e *blaster*, onde *Ensemble* e SVM também foram empregados, no gerenciamento de equipamentos, no gerenciamento geotécnico e na segurança de minas. Na etapa de recuperação ambiental destacaram-se os SVMs, usados principalmente para monitoramento da cobertura do solo e na avaliação de risco nas minas (JUNG; CHOI, 2021).

	EXPLORAÇÃO		EXTRAÇÃO (OPERAÇÃO)				RECUPERAÇÃO AMBIENTAL		
	Prospecção e Sondagem	Planejamento da mina	Perfuração e Blaster	Equipamentos	Geotécnica	Segurança	Monitoramento	Avaliação	TOTAL
DL	6	3	17	4	8	11	2	5	56
Ensemble	12	1	12	5	15	6	5	-	56
SVM	10	1	12	3	11	6	8	8	59
Bayesiano	-	-	-	-	3	3	1	2	9
AD	2	4	3	-	4	5	-	5	23
BI	1	-	3	-	2	3	1	1	11
Classificação	2	-	1	-	-	-	-	-	2
Regressão	4	-	11	1	2	1	-	4	23
RD	1	-	-	-	4	-	-	-	5
AG	-	-	-	1	1	-	-	-	2

Tabela 2: Número de trabalhos publicados usando os diferentes métodos de ML em cada etapa da mineração. DL: *Deep Learning*, AD: Árvore de Decisão, BI: Baseado em Instância, RD: Redução Dimensional, AG: Algoritmos Genéticos. **Fonte**: modificado de JUNG e CHOI (2021).

Por fim, os autores aplicaram métricas de avaliação para cada modelo de ML e observaram que os métodos com mais acurácia entre os trabalhos avaliados foram regressão, *Ensemble* e SVMs, respectivamente (Tabela 2) (JUNG; CHOI, 2021), todos modelos de aprendizado supervisionado/semi-supervisionado. É amplamente demonstrado na literatura que os modelos supervisionados tendem a ser mais precisos, pois cada um dos classificadores é treinado em uma coleção de dados representativos conhecidos, no entanto eles requerem maior quantidade de tempo para treinar os modelos, o que pode limitar sua aplicação em bancos de dados muito grandes (CHAOVALIT; ZHOU, 2005).

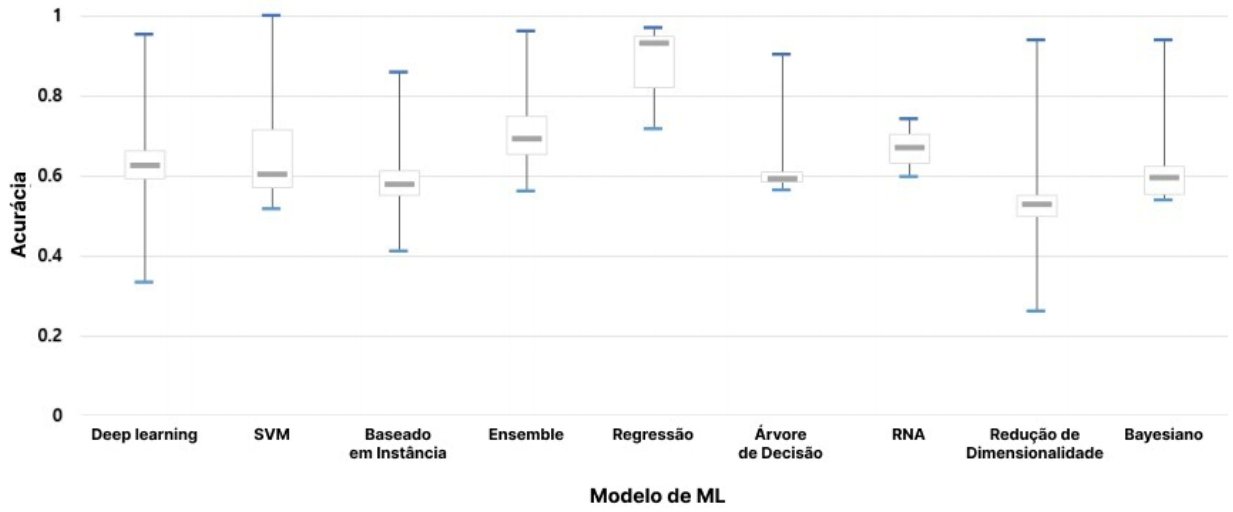


Figura 26: Acurácia dos diferentes modelos de ML usados na mineração. **Fonte**: modificado de JUNG e CHOI (2021).

3.4 O algoritmo *Random Forest*

Nas geociências existem várias análises que envolvem bancos de dados contendo tanto variáveis contínuas como variáveis categóricas no mesmo conjunto de dados, ou seja, requerem o uso de algoritmos tanto de classificação como regressão. Nesse tipo de base de dados o algoritmo de aprendizado supervisionado *Random Forest* (RF) é amplamente utilizado (SHENG *et al*., 2015; TOHRY *et al*., 2022).

O RF é um método não paramétrico multivariado, i.e., os dados da amostra não precisam seguir uma distribuição normal e múltiplas variáveis podem ser trabalhadas ao

mesmo tempo, de forma que é possível melhorar a precisão da classificação e determinar a importância das variáveis existentes em uma classificação (POLIKAR, 2012). O RF é um tipo de classificador *Ensemble* com comitê de aprendizes fracos, ou seja, com precisão um pouco melhor do que 50% (produzem um classificador que é apenas um pouco mais preciso do que a classificação aleatória), conforme explica Polikar (2012):

> Para classificação binária, é bem conhecido que o requisito exato para aprendizes fracos é ser melhor do que uma suposição aleatória. […] exigir que os aprendizes de base sejam melhores do que suposições aleatórias é muito fraco para problemas multiclasse, mas exigir mais de 50% de precisão é muito rigoroso (POLIKAR, 2012, p. 46).

Esse algoritmo se baseia no método estatístico de árvore de decisão, em que a classificação ou o valor de uma variável é predito com base em várias variáveis de entrada como discutido na sessão 3.2.4. Assim como no DL o poder de análise é amplificado combinando camadas mais profundas de nós/neurônios, o RF emprega árvores de decisão mais profundas durante o processo de treinamento (como indica o nome, é uma “floresta” de árvores) (POLIKAR, 2012).

Cada nó na árvore de decisão trabalha em um subconjunto aleatório de dados do *dataset* para calcular a saída. As árvores de decisão individuais são construídas para cada amostra usando o método estatístico de reamostragem (*bootstrapping*) com substituição (LEE; ULLAH; WANG, 2020). Cada árvore de decisão irá gerar uma saída e o resultado é calculado com base na votação majoritária para classificação ou média para regressão (Figura 27) (POLIKAR, 2012; SHAHABI *et al.*, 2019).

Como cada árvore de decisão em RF assume um voto para a classificação, a classe que obtém mais votos é escolhida como resultado do modelo (FAWAGREH; GABER; ELYAN, 2014). O uso de várias árvores no RF o torna menos propenso a erros em comparação com as árvores de decisão convencionais e o uso de amostragem aleatória diminui a correlação entre as árvores e o sobreajuste (condição em que o modelo se ajusta completamente aos dados de treinamento, mas não generaliza os dados de teste não vistos) (GHORBANZADEH; BLASCHKE, 2019; PARMAR; KATARIYA; PATEL, 2019).

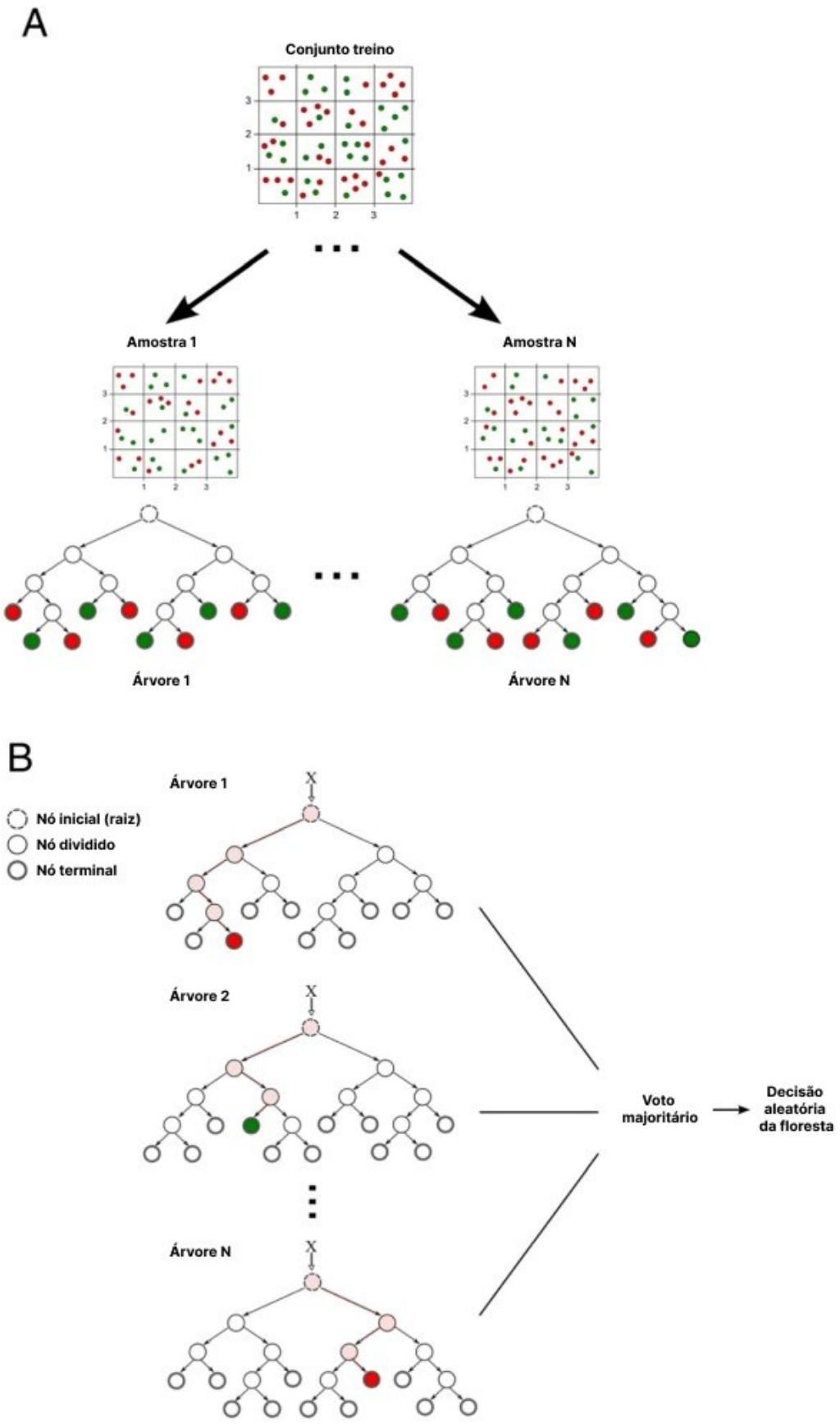

A
Conjunto treino
Amostra 1
Amostra N
Árvore 1
Árvore N
B
Nó inicial (raiz)
Nó dividido
Nó terminal
Árvore 1
Árvore 2
Árvore N
X
Voto majoritário
Decisão aleatória da floresta

Figura 27: Fases de treinamento e classificação do classificador RF. Círculos representam os neurônios e as linhas representam os galhos em cada árvore de decisão; neurônios de base correspondem à raiz e são representados por círculos tracejados, enquanto os neurônios terminais correspondem às folhas e são representados em negrito. **a)** Na fase de treinamento cada árvore de decisão no *ensemble* é construída sobre uma amostra aleatória de *bootstrap* dos dados originais, que contém exemplos positivos (rótulos verdes) e negativos (rótulos vermelhos). **b)** Na fase de classificação a previsão das classes é baseada em um procedimento de votação majoritária entre todas as árvores individuais. Em cada uma das árvores, para cada novo ponto de dados 'X', o algoritmo começa no neurônio raiz na base e percorre a árvore, testando os valores das variáveis em cada um dos neurônios divididos (rosa claro) e, para cada valor, seleciona o próximo ramo a seguir. Este processo é repetido até que um neurônio terminal no topo na árvore seja alcançado e atribua um rótulo de classe positiva ou negativa à categoria. No final do processo, cada árvore vota no rótulo de classe predito e a classe de saída final é predita. **Fonte:** MACHADO; MENDOZA; CORBELLINI (2015).

Recentemente, modelos RF foram empregados com sucesso em diferentes aplicações de mineração, especialmente em conjuntos de dados geoquímicos, pois observou-se que eles resultam em maior precisão do que outros modelos estatísticos convencionais e do que outros algoritmos de ML. O RF foi usado com sucesso para mapeamento geoquímico de Cu (RODRIGUEZ-GALIANO; CHICA-OLMO; CHICA-RIVAS, 2014), Cu-Au (KEYKHAY-HOSSEINPOOR *et al.*, 2020), Ag–Pb–Zn (WANG; ZHOU; XIAO, 2020) e prospecção mineral (RODRIGUEZ-GALIANO *et al*., 2015). Em comparação com RNA, SVM e árvores de regressão, foi mostrado que RF superou o resto dos métodos valores gerais de acurácia entre 92 e 96% (RODRIGUEZ-GALIANO *et al*., 2015).

Flores et al. (2019) compararam diversos algoritmo de ML e concluíram que o RF foi o mais eficiente em prever a recuperação de cobre por lixiviação coletados em campo pela empresa SCM Franke, no Chile. Foram obtidos valores bem classificados de 98,90% para dados operacionais e 98,72% para dados de pilha/estaca. Os autores também destacam que o uso deste algoritmo produziu resultados superiores as análises

previamente publicadas para recuperação de Cobre que se baseavam em modelos lineares (FLORES; KEITH; LEIVA, 2020).

No entanto, dependendo das características da base de dados, valores de acurácia um pouco menores também foram considerados efetivos em aplicações de mineração. Schnitzler e colaboradores (2019) empregaram RF para estimar os níveis de sódio a partir de dados químico-físicos medidos no depósito de sulfeto maciço vulcanogênico McLeod, no campo de mineração Matagami, Canadá. O uso de ML, nesse caso, retornou correlações de 0,66 a 0,75, dependendo dos conjuntos de treinamento e teste (SCHNITZLER; ROSS; GLOAGUEN, 2019).

Carranza e Laborte (2015a) usaram RF para prospecção mineral de depósitos de Ouro em Baguio (Filipinas) e mostraram que o RF consegue capturar com precisão e reprodutibilidade as relações espaciais entre as variáveis preditoras e os locais de depósito/não depósito de forma superior a métodos tradicionais de regressão logística (CARRANZA; LABORTE, 2015a).

4. RESULTADOS E DISCUSSÃO

4.1 Análise Exploratória dos Dados

A matriz de dispersão feita para a AED (Figura 28) mostrou que no *dataset* analisado, os dados da variável Fósforo (P%) não são confiáveis. É possível observar que as concentrações deste elemento são tão baixas que se pode considerar que o fósforo foi indetectável no *dataset* usado, sugerindo que o P% não tem correlação com os outros elementos. Dessa forma o P% é apenas um artefato e pode-se desconsiderar quaisquer relações com esta variável mesmo antes de se realizar a análise supervisionada. Também se observou que a variável relativa ao óxido de Ferro (FeO) apresentou correlação positiva e negativa com outras variáveis, mais notadamente com óxido de silício (SiO_2) e alumina (Al_2O_3) e parcialmente com MgO (Figura 28 e Anexo II), por exemplo.

Os gráficos de dispersão individuais para todos os elementos encontram-se no Anexo II e fornecem informações importantes no contexto da mineração, pois é possível checar, via correlação direta, como a concentração de um elemento ou processo afeta a recuperação mássica do Ferro, desde que esses sejam dados paramétricos, pois este é um dos requisitos para esse tipo de correlação.

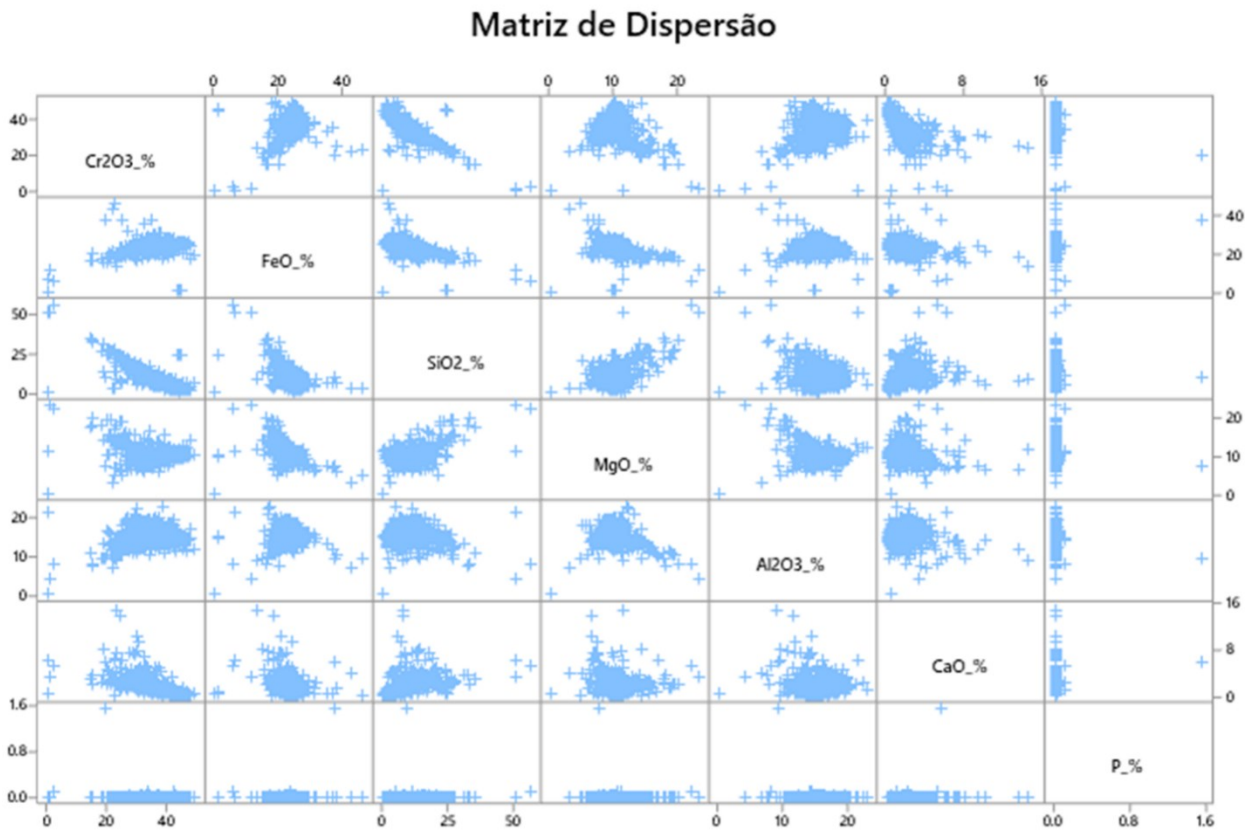


Figura 28: Matriz de dispersão mostrando a correlação dos elementos químicos presentes no *dataset*.

A análise de distribuição para os dados relativos à porcentagem de óxido de Ferro (FeO) no *dataset* analisado é apresentada na Figura 29. Nota-se que a ocorrência de FeO apresenta uma distribuição gaussiana normal para em torno de 24%. (PARMAR; KATARIYA; PATEL, 2019). A distribuição das concentrações dos demais elementos presentes no dataset é apresentada no Anexo III.

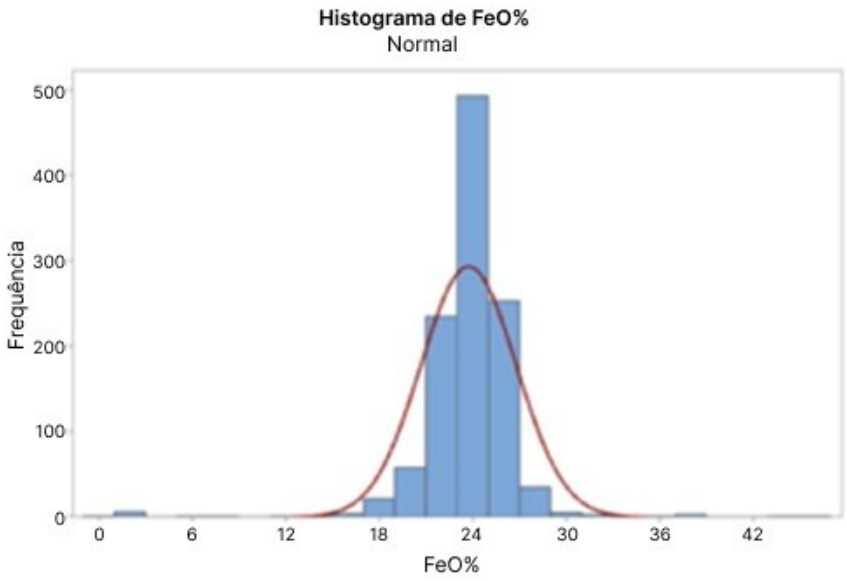


Figura 29: Histograma de distribuição da concentração de FeO no *dataset* analisado.

Através da análise por regressão pode-se notar um bom ajuste da distribuição dos dados para o FeO (Figura 30a), mas não para o P (Figura 30b). Isso está indicado na estatística de adequação de Anderson-Darling (AD), uma distância ao quadrado que mede a área entre a linha ajustada (com base na distribuição normal) e a função de distribuição empírica (com base nos pontos de dados). Valores maiores para a estatística de AD indicam que os dados não seguem a distribuição normal.

Em acordo com a informação mostrada no histograma da Figura 29, o FeO com padrão de distribuição normal no plano cartesiano tem valor AD = 51 e bom *fit* entre a distribuição dos dados (em azul) e a regressão teórica (em vermelho). Em contraste, o P não apresenta distribuição que se ajuste a um modelo linear, o que indica que há uma distribuição não normal no *dataset*, tal como observado na Figura 30b (AD = 392). Esta distribuição não normal pode estar associada à própria natureza do tipo de dado ou por problemas de detecção, erros e/ou dados faltantes.

As distribuições probabilísticas dos demais elementos constam no Anexo IV.

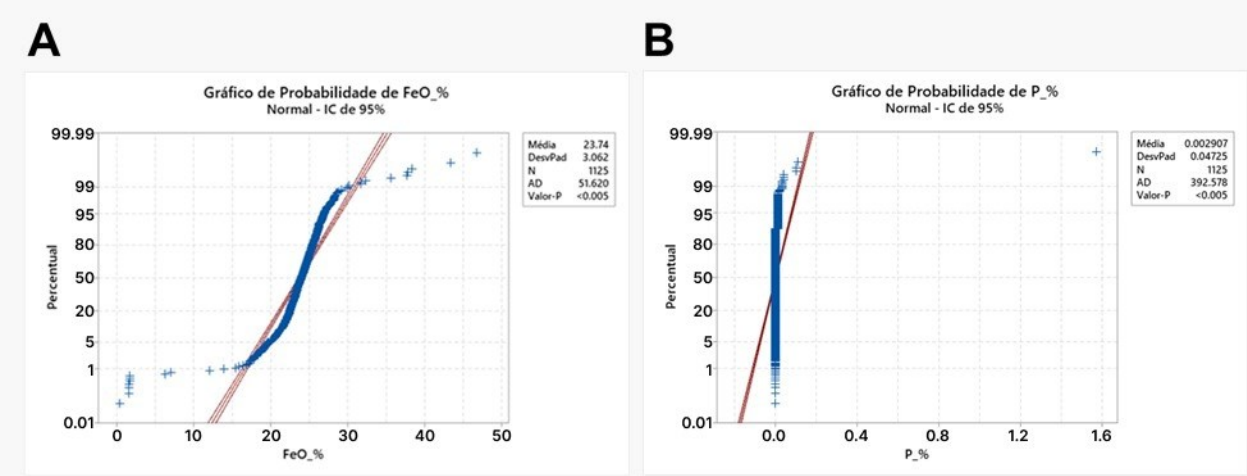


Figura 30: Gráficos de probabilidade indicando a distribuição (em azul) dos dados de FeO (**a**) e P (**b**) no plano cartesiano e seu ajuste a uma regressão linear teórica (em vermelho). DesvPad: desvio padrão; AD: adequação de Anderson-Darling.

Segmentações mais bem delimitadas da distribuição dos dados entre os quartis para a identificação de possíveis *outliers* são apresentados na Figura 31. Esta visualização permite inferir informações sobre distribuições antes mesmo de serem calculadas as normalidades ou distribuições de probabilidade. Quando a distribuição apresentada para um elemento é simétrica, a maioria dos dados está localizada ao redor da mediana, entre o primeiro e o terceiro quartil. Distribuições assimétricas indicam a possibilidade de que os dados não apresentem distribuição normal. Em concordância com o observado na distribuição probabilista (ver Figura 29) os dados possuem tendência de estarem distribuídos ao redor da média, com pequena distorção à esquerda para os teores de Cr_2O_3 e a direita para a SiO_2. Novamente podem ser observados valores muito baixos em toda distribuição do P, formando uma sobreposição entre todos os quartis.

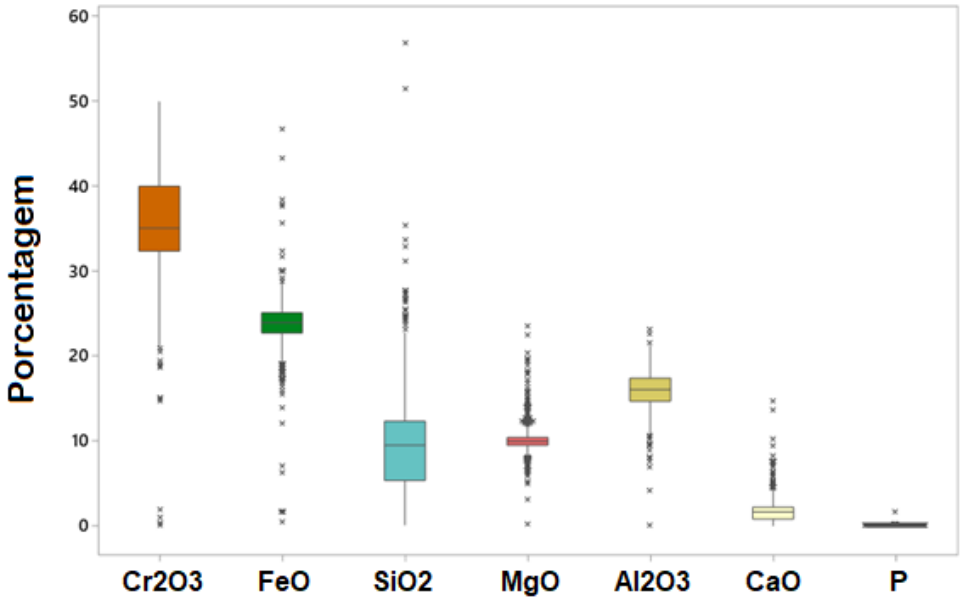


Figura 31: Distribuições referentes às concentrações de Cr_2O_3, FeO, SiO_2, MgO, Al_2O_3, CaO e P no *dataset* analisado.

4.2 Treinamento e teste do modelo empregando o *Random Forest*

A matriz gerada para o início do treinamento é apresentada na Figura 32. Nela, quanto mais vermelho, maior a relação das variáveis preditoras com altos índices de recuperação mássica de Ferro (classe 1), pois a classe **True** (valor numérico 1) foi atribuída quando o valor da concentração de Ferro no rejeito fosse alto (maior que o valor da média de Ferro entre todas as amostras do *dataset*) ou **False** (valor numérico 0) caso fosse baixa (menor que a média) (detalhes descritos na sessão 2.4).

Destaca-se que a escolha de corte a partir da média é um parâmetro arbitrário usado automaticamente pelo modelo, de acordo com a média de dados apenas do *dataset* usado neste trabalho, o que não necessariamente reflete o que é considerado como valor aceitável de Ferro recuperado de rejeito nas operações industriais de mineração. Assim, o modelo aqui apresentado funciona apenas como um exercício demonstrativo da aplicação do RF em bancos de dados públicos. Para a aplicação prática dessa estratégia em dados de campo ou durante as operações de mineração

basta modificar os parâmetros de porcentagem de Ferro no rejeito considerados adequados para recuperação mássica.

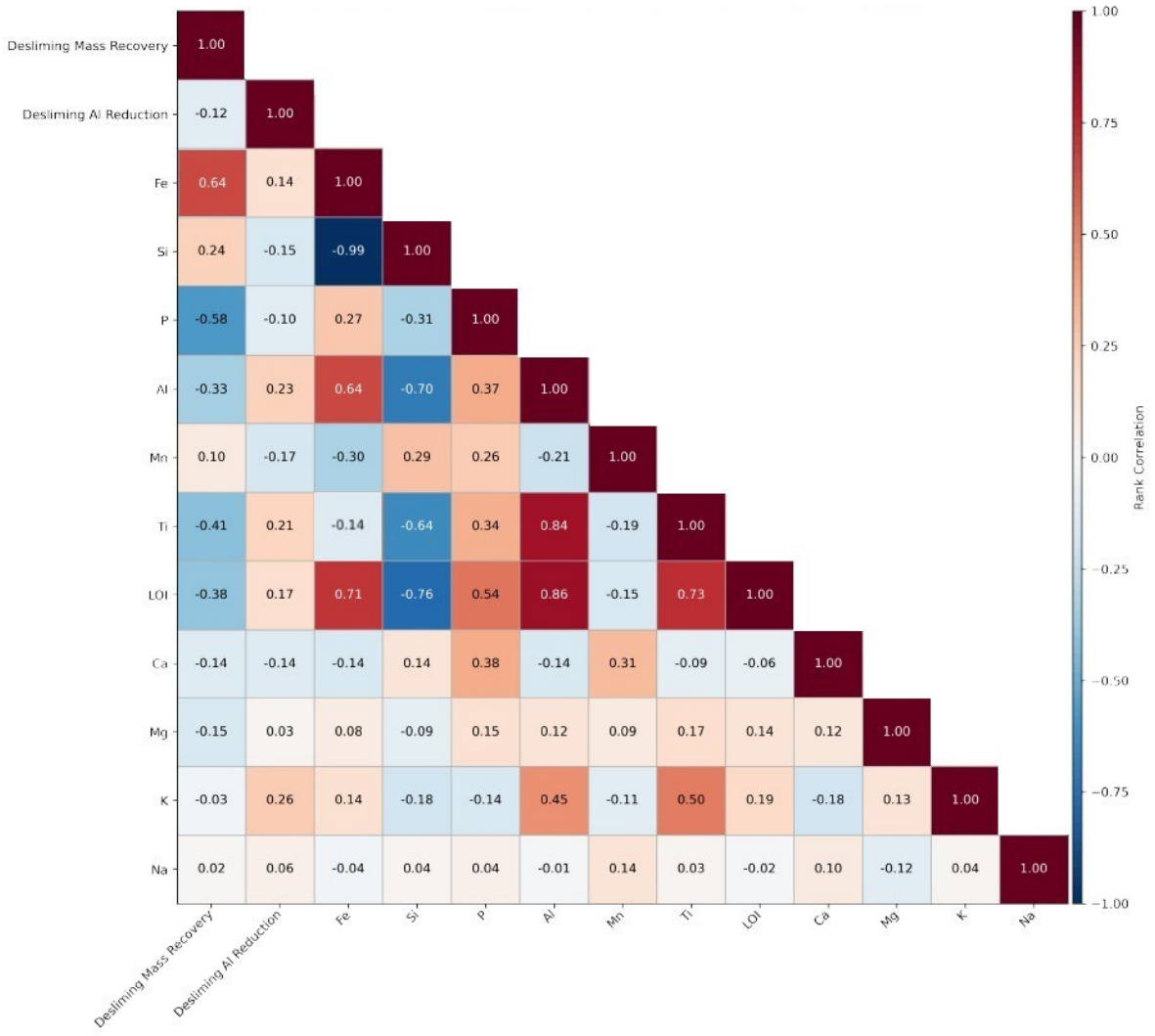


Figura 32: Matriz gerada após treinamento do RF. A matriz foi gerada usando os dados de porcentagem de Cr_2O_3, FeO, SiO_2, MgO, $Al2O_3$, CaO e P e relativos à porcentagem destes minerais após diferentes etapas do beneficiamento: deslamagem (*desliming mass recovery*), redução de alumina (*desliming Al reduction*) e perda por calcinação (LOI).

As métricas de performance do modelo obtido são apresentadas na Tabela 3. Nota-se que no modelo testado houve um desbalanço entre as classes, já que houve

mais que o dobro de variáveis preditoras na classe baixa recuperação mássica em relação a classe alta recuperação mássica. A acurácia geral obtida foi de 0,74, ou seja, o modelo classificou corretamente entre alta e baixa recuperação mássica de Ferro em 74% das vezes no conjunto testado. Sabendo que os índices relativos ao P não são confiáveis, como observado na AED, é possível que esse parâmetro tenha afetado a performance do modelo. Observa-se também um desequilíbrio entre as classes, visto que o valor de Suporte obtido na classe 0 (baixo Ferro) foi obtido mais do que o dobro do valor de Suporte obtido na classe 1 (alto Ferro). Como em um conjunto de dados desequilibrado o *F1-score* capturará melhor o desempenho do modelo em cada classe, tanto a precisão quanto a F1 estão refletindo um desempenho geral mais alto na categoria 0 do que na categoria 1.

Tabela 3: Métricas de performance do modelo gerado com RF

	Precisão	***Recall***	**F1-Score**	**Suporte**
1	0,67	0,69	0,62	73
0	0,80	0,80	0,82	151
Acurácia	-	-	**0,74**	231
Macro AVG	0,73	0,74	0,72	231
Peso AVG	0,76	0,77	0,74	231

4.3 Discussão

A análise exploratória dos dados foi uma estratégia bem-sucedida no fornecimento de percepções sobre os dados, que ajudaram na elaboração de hipóteses e na discriminação da natureza das distribuições e correlações entre parâmetros. Por exemplo, foi possível confirmar dados previamente publicados e ilustrados na Tabela 1 da correlação negativa entre a maior presença de quartzo (representado por SiO_2) com menores concentrações de Ferro. Esse tipo de informação, somada a outros achados, permite levantar hipóteses sobre o tipo de minério de Ferro presente num conjunto de dados e a natureza da formação geológica. Em amostras do tipo Carajás, por exemplo,

essa relação ocorre nas rochas dessa formação, em especial na região amazônica, a quantidade de quartzo diminui perto da superfície, enquanto a quantidade de Ferro aumenta em direção ao topo (DA SILVA; DA COSTA, 2020).

Outra relação importante e conhecida, para algumas amostras de furo de sondagem, são as concentrações de P inversamente relacionadas a maiores concentrações de Ferro (SCHUNNESSON, 1990), no entanto neste *dataset* este preditor de classe foi impreciso e pode ter a acurácia do modelo gerado. O emprego da AED permitiu a detecção da falta de confiabilidade dos dados deste elemento, o que ajuda a levantar a hipótese que adicionar outro conjunto de dados contendo dados mais robustos para o P ou mesmo omitir esse preditor na classificação do RF podem aumentar a acurácia do modelo.

Estes *insights* são importantes tanto no entendimento dos dados, como da natureza da amostra e são compatíveis com achados da literatura que comumente empregam modelos estatísticos tradicionais para predição de minerais como o Ferro. Por exemplo, está bem documentado na literatura a correlação entre os teores de sílica e a recuperação mássica de Ferro durante o beneficiamento (ARANTES; LIMA, 2013; MARTINS *et al.*, 2017). No entanto, é notável que mesmo em relações estabelecidas, como a relação inversamente proporcional entre os teores de Fe_2O_3 e os teores de SiO_2 existem materiais, especialmente rejeitos, onde outros fatores irão afetar o teor de Ferro no rejeito independentemente dos teores de SiO_2 (Tabela 1). Ou seja, mesmo em parâmetros fortemente correlacionados, as análises estatísticas convencionais são limitadas para criar modelos que possam explicar uma variável (como recuperação mássica) em relação a dezenas de outras variáveis sabidamente importantes tanto na composição do mineral no sítio, como nos processos de beneficiamento.

Nesse sentido, esse estudo consegue combinar múltiplas variáveis que são analisadas concomitantemente, para criar um sistema de predição de recuperação de Ferro através de ML, uma aplicação inédita para esse mineral e que leva em conta tanto parâmetros geoquímicos (concentrações de elementos tipicamente co-ocorrentes) como o efeito sobre todos estes dos processos de beneficiamento, uma vez que há parâmetros, como as concentrações de Alumina, que afetarão a recuperação mássica

durante o beneficiamento industrial (DONG *et al.*, 2014; RAGHUKUMAR; KUMAR TRIPATHY; MOHANAN, 2012).

O modelo apresentado neste trabalho usou 42 árvores e duas classes para a classificação de um tipo de minério. A quantidade de árvores, que são os neurônios de processamento, usada na análise é um parâmetro que influencia a acurácia do RF. Em comparação, Sheng *et al.* (2015) apresentaram um trabalho visando a identificação e classificação da qualidade de minério de Ferro, usando como padrão de determinação da qualidade do minério de Ferro amostras industriais classificadas em 10 classes de qualidade para treinamento de um modelo de RF. Os autores realizaram múltiplas etapas de otimização para obter um modelo RF com acurácia de 100% de identificação e classificação dos 10 tipos de minério de Ferro. Nesse contexto, foi observada a necessidade de uso de 72 árvores e 300 variáveis aleatórias.

Por outro lado, Rodriguez-Galiano *et al.* (2015) mostraram que valores muito pequenos do número de árvores resultaram em desempenho de previsão inferior, de forma que para trabalhos futuros se recomenda aumentar o número de árvores e variáveis classificadoras usadas no modelo de classificação para recuperação mássica de Ferro como estratégia para obter melhor acurácia geral, mas destaca-se que isso aumentará o custo computacional e tempo de processamento dos dados.

Uma vez que existem diversos graus de minério de Ferro (SHENG *et al*., 2015) e rejeitos de várias naturezas e diferentes composições (Tabela 1), a reclassificação das classes de acordo com porcentagem de Ferro próxima das concentrações consideradas adequadas para recuperação mássica em operações de mineração ou de reaproveitamento de rejeitos também é uma estratégia que poderá gerar melhor distribuição entre as classes aqui empregadas, permitindo também a aplicação prática deste modelo.

Uma consequência da criação de duas classes divididas arbitrariamente com base na média de Ferro no *dataset* foi a sub-representação de classificadores de alto Ferro (classe 1 = 73) em relação a classificadores de baixo Ferro (classe 0 = 151), como pode ser observado no parâmetro suporte da Tabela 3. Como o RF é otimizado para criar modelos de discriminação que têm o maior sucesso geral de classificação, as classes que incluem mais observações geralmente são mais bem classificadas em comparação

aquelas em que menos observações estão disponíveis (CHEN; LIAW, 2004). Num caso similar (SCHÖNIG *et al.*, 2021) aplicaram o modelo de discriminação *garnet* sobre o RF para equilibrar as taxas de sucesso de classificação para as classes individuais da melhor forma possível, o que melhorou o *Recall* da classe sub-representada, aumentando a acurácia desta classe. Essa estratégia pode ser particularmente interessante uma vez que melhoraria o poder preditivo sobre a classe que revela maior recuperação mássica de Ferro no *dataset* aqui empregado.

No presente trabalho foram mantidas todas as 13 variáveis preditoras, incluindo o P, como verificado em outros estudos (GHIMIRE et al., 2011; SHENG *et al*., 2015a; ZHANG *et al*., 2016), no entanto Carranza e Laborte (2015b) mostraram que o uso apenas dos melhores preditores produziu maior acurácia geral do modelo de prospecção de Cobre em Abra (Filipinas). Ao usar todos os preditores foi observada acurácia de 0,67 e usando apenas os três preditores mais importantes e significativos foi de 0,75 (CARRANZA; LABORTE, 2015b). Essa estratégia também pode ser empregada em trabalhos futuros para testar o efeito sobre a acurácia do modelo de recuperação mássica de Ferro.

Muitos campos da ciência quantitativa estão adotando o paradigma da modelagem preditiva usando ML, no entanto amplas revisões, em trabalhos usando ML nas múltiplas áreas do conhecimento, levantam o alerta para problemas de reprodutibilidade computacional, que seria quando os resultados de um trabalho podem ser replicados usando o código exato e o conjunto de dados fornecidos pelos autores (KAPOOR; NARAYANAN, 2022). Como já citado anteriormente, o uso do código e parâmetro usados neste trabalho deve ser avaliado com cautela antes de ser aplicado em outros conjuntos de dados. É preciso que parâmetros de treinamento e de critérios de classificação, em especial as concentrações de Ferro usadas como parâmetro classificador, sejam ajustados de acordo com os objetivos específicos desejado e características dos dados usados.

Essa abordagem abre caminho para cruzar estes parâmetros de classificação de minério de Ferro em múltiplas classes de qualidade, não apenas duas como usado neste trabalho, além disso, sugere que pode ser possível aumentar a acurácia do modelo aqui gerado usando mais árvores durante a etapa de treinamento, assim como a possibilidade

de combinar múltiplos *datasets* com parâmetros semelhantes e relacionados a mineração e beneficiamento de Ferro, para obter tanto mais parâmetros classificadores, como aumentar o ‘n’ total do conjunto de dados, o que torna os modelos mais robustos e precisos, como já mostrado anteriormente para o RF (GHIMIRE *et al.*, 2011; PROBST; BOULESTEIX, 2018; VANSCHOREN; BLOCKEEL, 2010).

Por fim, enquanto modelos estatísticos convencionais, como análises de regressão, são frequentemente empregados na modelagem de prospecção mineral, pois elas permitem obter conhecimento sobre o fenômeno subjacente, eles são limitados quando se deseja entender a relação entre muitas variáveis. Por outro lado, técnicas de ML, como RF, cumprem bem o papel de realizar previsões em conjuntos de dados complexos, mas eles normalmente fornecem pouca visão sobre o problema em relação à importância ou relevância das covariáveis individuais (KOST; RHEINBACH; SCHAEBEN, 2021). Historicamente essas aplicações foram empregadas na estatística e geoestatística de forma separada (BREIMAN, 2003), no entanto, a evolução dos métodos computacionais e ferramentas de ciência de dados permite uma combinação, em que são realizadas 1) explorações estatísticas convencionais, como na AED, com objetivo de entendimento dos dados e da relação direta entre variáveis, onde a precisão preditiva é secundária e 2) o uso de algoritmos de ML, em que a precisão preditiva é o objetivo. Isso permite o entendimento de um problema de forma complementar e mais adequada aos desafios complexos que existem nas Geociências e mineração modernas.

5. CONCLUSÃO

A partir da revisão da literatura e das análises apresentadas, pode-se concluir que os métodos de ML são uma ferramenta importante e com alta eficiência, que vem sendo cada vez mais utilizada em todas as fases da mineração. Há neste âmbito um amplo uso do algoritmo *Random Forest*, especialmente em problemas de classificação com bancos de dados complexos, tanto os disponíveis em bancos públicos, como o usado neste trabalho, quanto dados de campo e de sensores terrestres ou aéreos.

As análises feitas a partir de um banco com dados de teores de diversos minerais, incluindo o óxido de Ferro, se mostrou um método rápido, sem custo e eficiente que pode fornecer várias informações importantes. A análise exploratória dos dados foi uma estratégia bem-sucedida no fornecimento de percepções sobre os dados, que ajudaram na elaboração de hipóteses e na discriminação da natureza das distribuições e correlações entre parâmetros.

A aplicação do algoritmo de aprendizagem supervisionada *Random Forest* foi eficaz em criar um modelo para indicar alta ou baixa recuperação mássica de Ferro, uma aplicação inédita, com acurácia geral de 74%, a partir de 13 variáveis lito-geoquímicas contendo informações dos teores dos principais minerais co-ocorrentes em minérios de Ferro antes e após algumas das principais etapas de beneficiamento do Ferro na mineração brasileira.

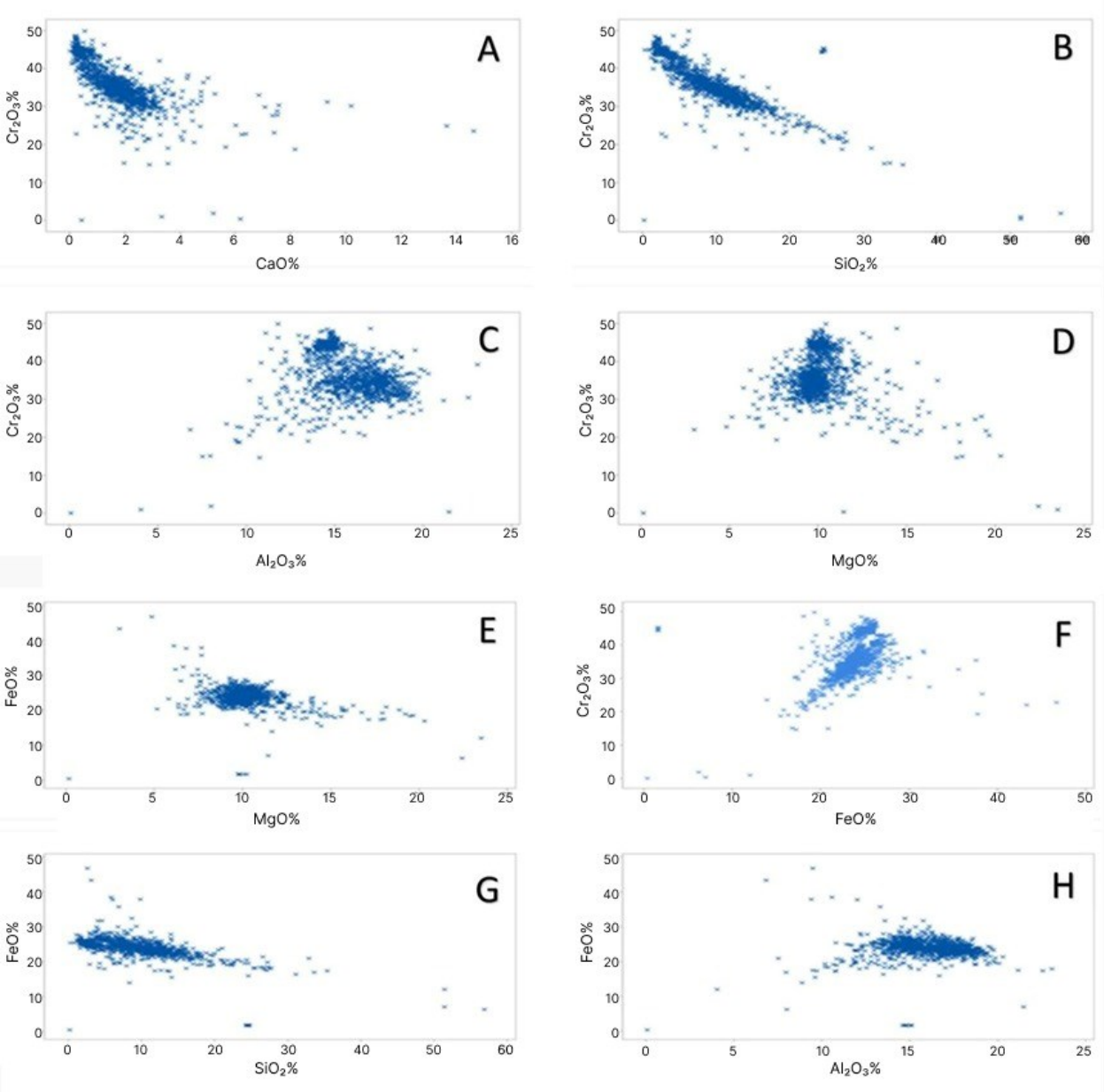
Para fins de aplicações práticas em trabalhos futuros, sugere-se adequar o número de classes de grau de Ferro e suas respectivas concentrações de acordo com o tipo de amostra contendo Ferro a ser avaliada, além de ajustes no número de arvores, combinação com outros *datasets* ou seleção do número de variáveis preditoras de acordo com a capacidade computacional disponível e escopo da pesquisa, o que pode possibilitar maior acurácia.

ANEXO I: CÓDIGO EM PYTHON USADO NA AED E RF

```
# Importação dos pacotes
import pandas as pd
import matplotlib as mat
import matplotlib.pyplot as plt
import numpy as np
df = pd.read_csv("geology_lito.csv")
# Verificação da existência de valores Nulos
df.isnull().values.any()
# Identificação de correlação das variáveis
def plot_corr(df, size=10):
    corr = df.corr()
    fig, ax = plt.subplots(figsize = (size, size))
    ax.matshow(corr)
    plt.xticks(range(len(corr.columns)), corr.columns)
    plt.yticks(range(len(corr.columns)), corr.columns)
df.corr()
# Definição das classes
ferej_map = {True : 1, False : 0}
# Aplicação de mapeamento no dataframe
df['ferej_map'] = df['ferej_map'].map(ferej_map)
# Verificação na distribuição dos dados
num_true = len(df.loc[df['ferej_map'] == True])
num_false = len(df.loc[df['ferej_map'] == False])
#Treinamento da % dos dados
from sklearn.model_selection import train_test_split
# Seleção de variáveis preditoras (Feature Selection)
atributos = ['fe', 'al', 'mgo', 'fe_deslam', 'fe_rej', 'sio2', 'p', 'mn','ti', 'loi', 'ca', 'na','k']
atrib_prev = ['rec_massiv']
X = df[atributos].values
Y = df[atrib_prev].values
# Taxa de split dos dados
split_test_size = 0.30
# Gerando os dados de Treino e teste
X_treino, X_teste, Y_treino, Y_teste = train_test_split(X, Y, test_size = split_test_size, random_state = 42)
print("{0:0.2f}% nos dados de treino".format((len(X_treino)/len(df.index)) * 100))
print("{0:0.2f}% nos dados de teste".format((len(X_teste)/len(df.index)) * 100))
print("Original True : {0} ({1:0.2f}%)".format(len(df.loc[df['ferej'] == 1]),
                                        (len(df.loc[df['ferej'] ==1])/len(df.index) * 100)))

print("Original False : {0} ({1:0.2f}%)".format(len(df.loc[df['ferej'] == 0]),
```

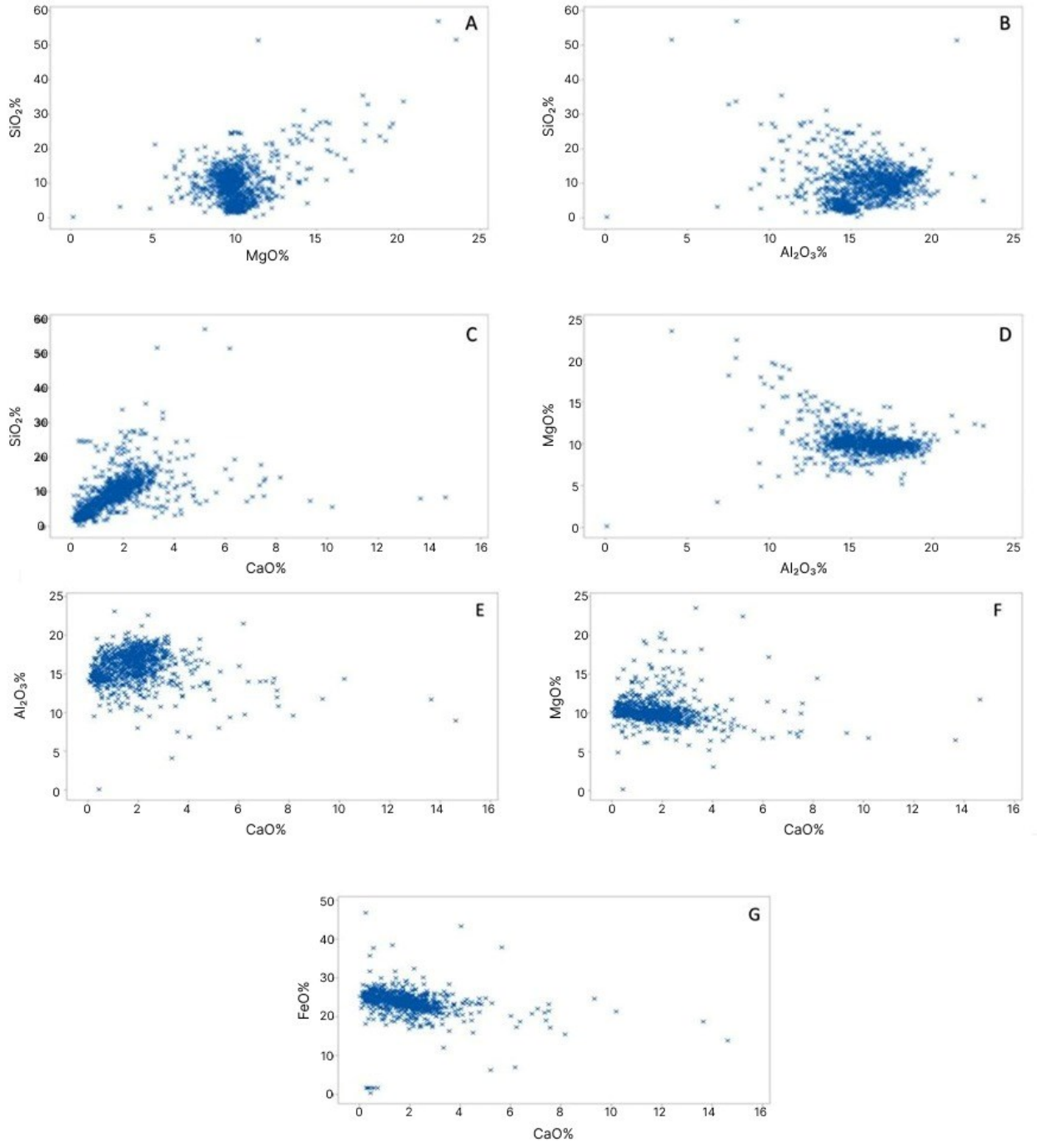
```
                                    (len(df.loc[df['ferej'] == 0])/len(df.index) * 100)))
print("")
print("Training True : {0} ({1:0.2f}%)".format(len(Y_treino[Y_treino[:] == 1]),
                                    (len(Y_treino[Y_treino[:] == 1])/len(Y_treino) * 100)))

print("Training False : {0} ({1:0.2f}%)".format(len(Y_treino[Y_treino[:] == 0]),
                                    (len(Y_treino[Y_treino[:] == 0])/len(Y_treino) * 100)))
print("")
print("Test True : {0} ({1:0.2f}%)".format(len(Y_teste[Y_teste[:] == 1]),
                                    (len(Y_teste[Y_teste[:] == 1])/len(Y_teste) * 100)))

print("Test False : {0} ({1:0.2f}%)".format(len(Y_teste[Y_teste[:] == 0]),
                                    (len(Y_teste[Y_teste[:] == 0])/len(Y_teste) * 100)))

#Tratamento de dados faltantes
preenche_0 = SimpleImputer(missing_values = 0, strategy = "mean")
X_treino = preenche_0.fit_transform(X_treino)
X_teste = preenche_0.fit_transform(X_teste)

# Criar Matrix de confusão
print("{0}".format(metrics.confusion_matrix(Y_teste, nb_predict_test, labels = [1, 0])))
print("")
print("Classification Report")
print(metrics.classification_report(Y_teste, nb_predict_test, labels = [1, 0]))

#Otimizando o Modelo com RandomForest
from sklearn.ensemble import RandomForestClassifier
modelo_v2 = RandomForestClassifier(random_state = 42)
modelo_v2.fit(X_treino, Y_treino.ravel())
RandomForestClassifier(random_state=42)
# Verificando os dados de treino
rf_predict_train = modelo_v2.predict(X_treino)
print("Exatidão (Accuracy): {0:.4f}".format(metrics.accuracy_score(Y_treino, rf_predict_train)))
# Verificando nos dados de teste
rf_predict_test = modelo_v2.predict(X_teste)
print("Exatidão (Accuracy): {0:.4f}".format(metrics.accuracy_score(Y_teste, rf_predict_test)))
print()
```

ANEXO II: *Scatterplots* individuais de comparações usadas na matriz de dispersão



Scatterplots 1: Dispersões das comparações entre elementos.
A) O diagrama de dispersão mostra uma correlação negativa forte para baixos valores de CaO% (< 2%), e correlação negativa fraca para 2 < CaO% < 4. Para valores de CaO > 4 os valores mais dispersos indicam que não há boa correlação nessa faixa.
B) Correlação negativa forte entre Cr2O3% conforme aumentam os valores de SiO2%.
C) Os dados se agrupam na faixa entre 30 a 45 (Cr2O3%) e 15 e 20 (SiO2%) com uma tendência positiva muito fraca.
D) A maior parte dos dados se agrupam em valores altos de Cr2O3% e valores médios de MgO%.
E) Dados agrupados num cluster na faixa de 20 < FeO% < 30 que tende a ser constante com o aumento de MgO%.

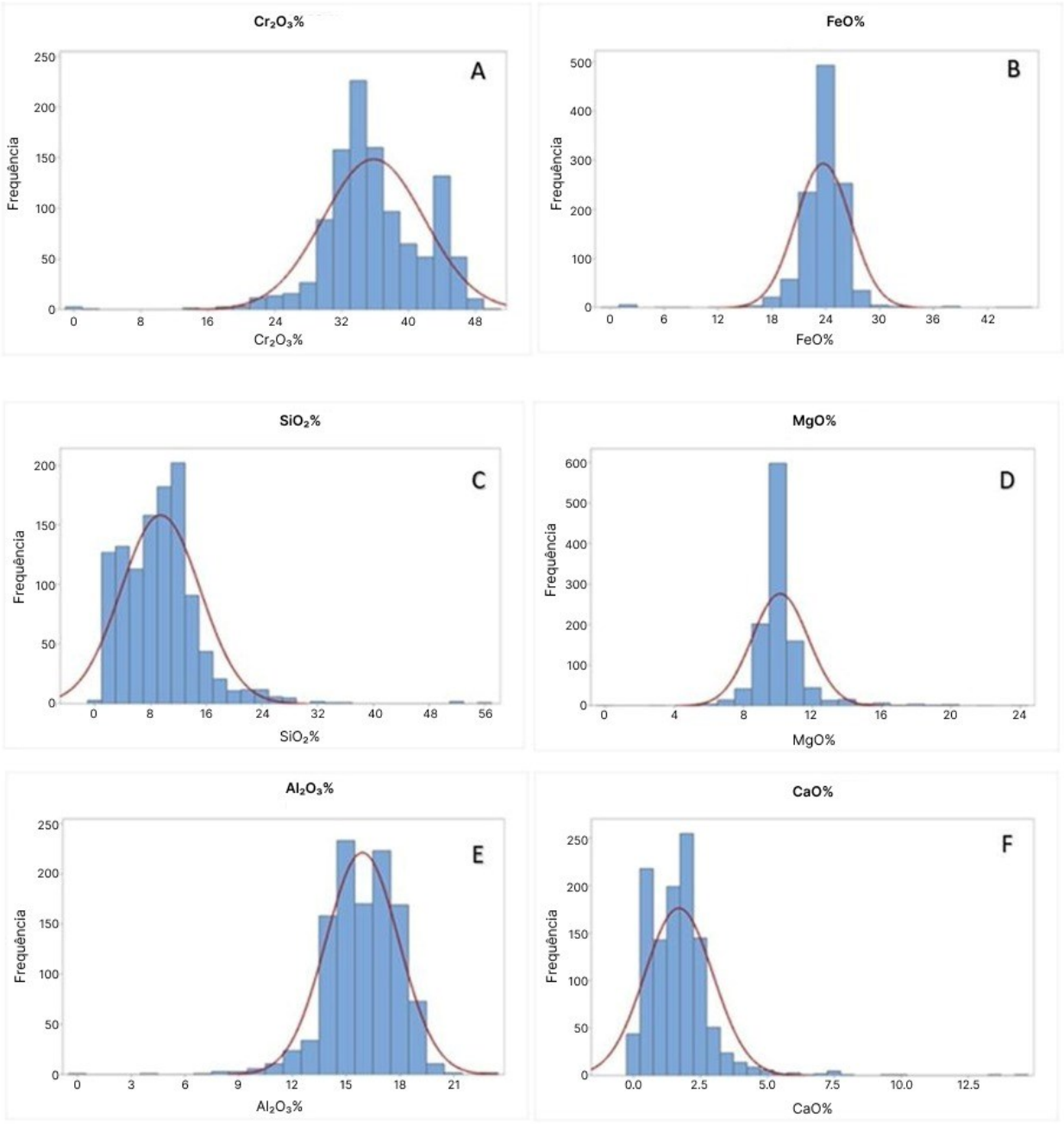
F) Correlação positiva forte de Cr2O3% com alguns dados dispersos.
G) Correlação negativa discreta quase constante de FeO% com o aumento de SiO2.
H) Apresenta um cluster constante para FeO% em altos valores de Al2O3% (>15%).



Scatterplots 2: Dispersões das comparações entre elementos (continuação).
A) Apresenta valores baixos de SiO2% se apresentando em cluster com tendência de correlação positiva muito baixa para valores de MgO% > 10.
B) Não apresenta tendência de correlação entre os dados.

C) Os dados apresentam uma tendência de correlação positiva alta até CaO% = 3, depois se dispersam.
D) Os valores de MgO% se apresentam com tendência constante nos valores de Al2O3% entre 15 e 20%.
E) Os dados se apresentam em cluster sem tendência de correlação.
F) Os dados possuem correlação negativa alta com tendência a se dispersar em valores maiores de CaO% e de MgO%.
G) Correlação observada entre e FeO% entre 20 e 30 e CaO% <6.

ANEXO III: Distribuição dos dados e linha de normalidade para os elementos $Cr2O_3$, FeO, SiO_2, MgO, Al_2O_3 e CaO.



Histogramas e avaliação de distribuição para cada um dos elementos avaliados.

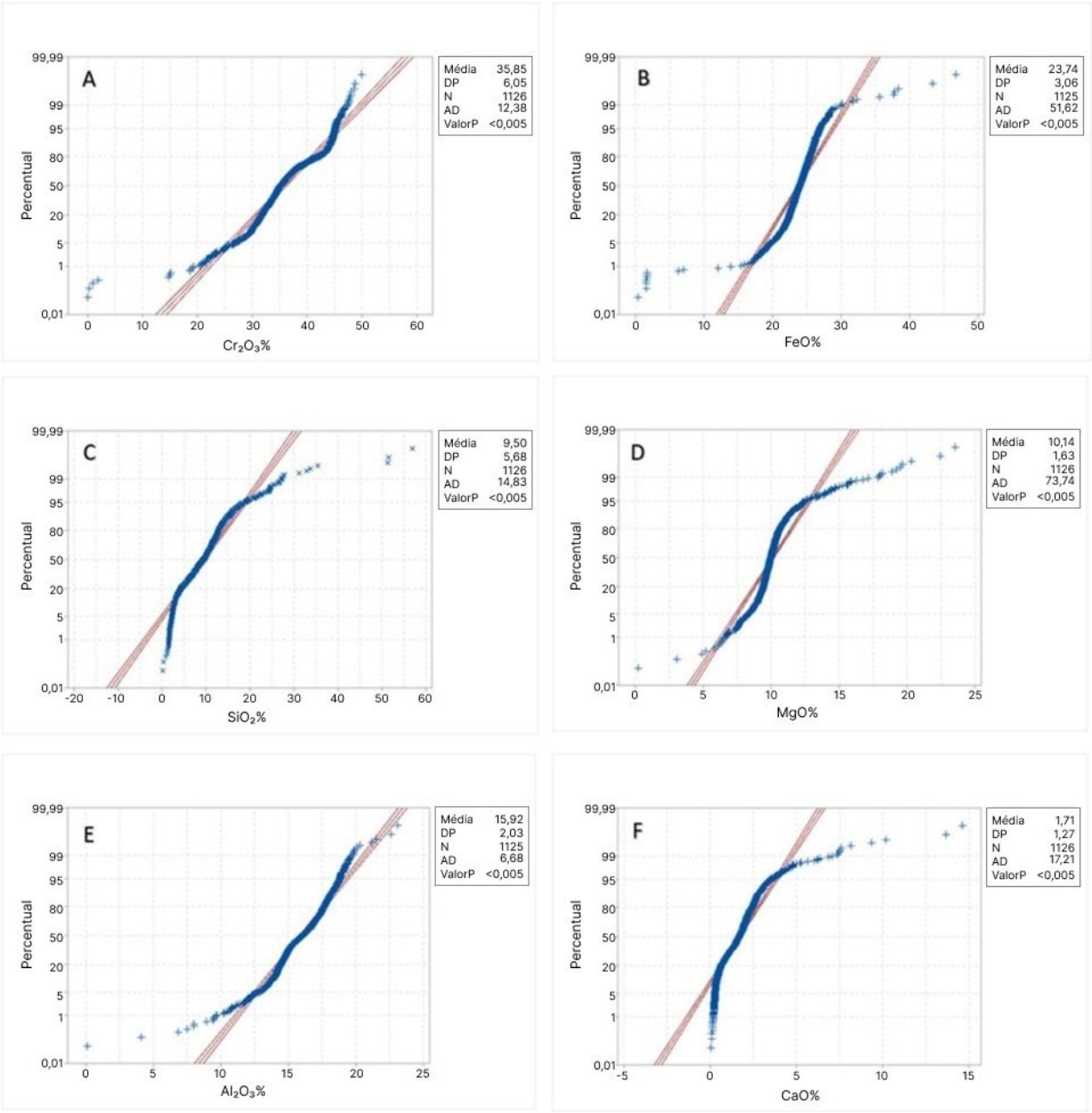
A) Diagrama mostra bimodalidade dos dados, com alta frequência de Cr2O3% nos valores de 34 e 44, onde a média simétrica dos dados se encontra no valor de 34% para frequência = 150.

B) O diagrama apresenta alta frequência em 23 FeO% e média simétrica em 10% para frequência no valor de 300.

C) Média simétrica na frequência de 150 para valores de SiO2% = 8.

D) Média simétrica no valor de MgO% = 10.
E) Média simétrica no valor de Al2O3 = 16, mostrando alta frequência dos dados.
F) Média simétrica no valor de CaO% = 1,5.

ANEXO IV: Distribuições de probabilidade referentes aos dados de $Cr2O_3$, FeO, SiO_2, MgO, Al_2O_3 e CaO.



Distribuição de probabilidade e linha de regressão teórica para cada um dos elementos avaliados.

A) O gráfico mostra uma normalidade na distribuição dos dados, apresentando muito pouco desvio para a distribuição.

B) Os dados não apresentam uma certa normalidade em sua distribuição média, sendo possível haver desvio, semelhantes com uma forma logarítmica nos teores de FeO% dos dados coletados.

C) A forma dos dados plotados indica que há um certo desvio na normalidade da distribuição, numa tendência de mudança para os valores mais altos de SiO2%.

D) Há um certo desvio na normalidade dos dados, sendo pouco para os valores de MgO% abaixo da média percentual, e é maior para os valores acima da média.
E) No geral há uma normalidade na distribuição dos dados, com pouca variação entre os dados de menor valor de Al2O3%.
F) Observa-se que há uma boa quantidade de percentual de dados com valores nulos e/ou próximos de nulos em CaO%, que não seguem uma normalidade na distribuição dos dados, e isso só volta a se repetir em valores CaO%>5. Na porção média dos valores de CaO% há um pequeno desvio, mas que não interfere na distribuição em geral.

REFERÊNCIAS BIBLIOGRÁFICAS

ACOSTA, I. C. C.; KHODADADZADEH, M.; TUSA, L.; GHAMISI, P.; GLOAGUEN, R. A Machine learning framework for drill-core mineral mapping using hyperspectral and high-resolution mineralogical data fusion. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 12, n. 12, p. 4829–4842, 2019.

AIZENBERG, Igor; AIZENBERG, Naum N.; VANDEWALLE, Joos PL. **Multi-valued and universal binary neurons: Theory, learning and applications**. Springer Science & Business Media, 2000.

ALAVI, A. H.; GANDOMI, A. H.; LARY, D. J. Progress of machine learning in geosciences: Preface. **Geoscience Frontiers**, v. 7, n. 1, p. 1–2, 2016.

ALLOGHANI, M.; AL-JUMEILY, D.; MUSTAFINA, J.; HUSSAIN, A.; ALJAAF, A. J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. **Supervised and unsupervised learning for data science**, p. 3-21, 2020.

ANGELOV, P. P.; SOARES, E. A.; JIANG, R.; ARNOLD, N. I.; ATKINSON, P. M. Explainable artificial intelligence: an analytical review. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 11, n. 5, p. e1424, 2021.

ARANTES, R. S.; LIMA, R. M. F. Influence of sodium silicate modulus on iron ore flotation with sodium oleate. **International Journal of Mineral Processing**, v. 125, p. 157–160, 2013.

BADEDA, J.; HUCK, M.; SAUER, D. U.; KABZINSKI, J.; WIRTH, J. Basics of lead–acid battery modelling and simulation. **Lead-Acid Batteries for Future Automobiles**, p. 463–507, 2017.

BARNARD, G. A.; WALD, A. Statistical Decision Functions. **Biometrika**, v. 40, n. 3/4, p. 475, dez. 1953.

BATISTA, J.; ANA, P.; PENHA, C.; ANA, P. Innovation policy for the mining sector in Brazil: a comparative analysis with Sweden centered on the interactions of involved agents. **Cadernos EBAPE.BR**, v. 17, n. 4, p. 959–974, 2019.

BEHRENS, J. T. Principles and Procedures of Exploratory Data Analysis. **Psychological Methods**, v. 2, n. 2, p. 131–160, 1997.

BENGIO, Y.; LOURADOUR, J.; COLLOBERT, R.; WESTON, J. Curriculum learning. **ACM International Conference Proceeding Series**, v. 382, 2009.

BERETTA, F.; RODRIGUES, A. L.; PERONI, R. L.; COSTA, J. F. C. L. Automated lithological classification using UAV and machine learning on an open cast mine. **Applied Earth Science**, v. 128, n. 3, p. 79-88, 2019.

BÉRUBÉ, C. L.; OLIVO, G. R.; CHOUTEAU, M.; PERROUTY, S.; SHAMSIPOUR, P.; ENKIN, R. J.; MORRIS, W. A.; FELTRIN, L.; THIÉMONGE, R. Predicting rock type and detecting hydrothermal alteration using machine learning and petrophysical properties of the Canadian Malartic ore and host rocks, Pontiac Subprovince, Québec, Canada. **Ore Geology Reviews**, v. 96, p. 130–145, 2018.

BRAGA, A. de P.; CARVALHO, A. C. P. de L. Ferreira.; LUDERMIR, T. Bernarda. **Redes neurais artificiais: teoria e aplicações**. LTC Editora, 2007. 226 p.

BRASIL. **Anuário Mineral Brasileiro: principais substâncias metálicas**. [s.l: s.n.]. Disponível em: <www.anm.gov.br>. Acesso em: 24 jul. 2022.

BREIMAN, L. Statistical modeling: The two cultures. **Quality control and applied statistics**, v. 48, n. 1, p. 81–82, 2003.

BRITISH COLUMBIA RGS. **Regional Geochemical Survey Data - Open Government Portal**. Disponível em: <https://open.canada.ca/data/en/dataset/49a73c36-4a25-4be5-b6d1-abd020fb031a>. Acesso em: 16 jul. 2022.

CARMIGNANO, O. R. D. R. **Inovação no setor de mineração de Ferro em Minas Gerais com foco na destinação de rejeitos**. 2021. Universidade Federal de Minas Gerais, Belo Horizonte, 2021. Disponível em: <https://repositorio.ufmg.br/bitstream/1843/38069/1/TESE.pdf>. Acesso em: 22 jul. 2022.

CARMIGNANO, O. R.; VIEIRA, S. S.; TEIXEIRA, A. P. C.; LAMEIRAS, F. S.; BRANDÃO, P. R. G.; LAGO, R. M. Iron Ore Tailings: Characterization and Applications. **Journal of the Brazilian Chemical Society**, v. 32, n. 10, p. 1895–1911, 2021.

CARPENTER, K. A.; COHEN, D. S.; JARRELL, J. T.; HUANG, X. Deep learning and virtual drug screening. **Future Medicinal Chemistry**, v. 10, n. 21, p. 2557–2567, 2018.

CARRANZA, E. J. M.; LABORTE, A. G. Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm. **Ore Geology Reviews**, v. 71, p. 777–787, 2015a.

CARRANZA, E. J. M.; LABORTE, A. G. Random Forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). **Computers & Geosciences**, v. 74, p. 60–70, 2015b.

CATÉ, A.; PEROZZI, L.; GLOAGUEN, E.; BLOUIN, M. Machine learning as a tool for geologists. **Leading Edge**, v. 36, n. 3, p. 215–219, 2017.

CERVANTES, J.; GARCIA-LAMONT, F.; RODRÍGUEZ-MAZAHUA, L.; LOPEZ, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. **Neurocomputing**, v. 408, p. 189–215, 2020.

CHAKRABORTI, S. **Wipro Mining Analytics. Big Data Analytics In Mining Industry**. 2022. Disponível em: <https://www.wipro.com/natural-resources/application-of-big-data-solution-to-mining-analytics/>. Acesso em: 6 jul. 2022.

CHAOVALIT, P.; ZHOU, L. Movie review mining: A comparison between supervised and unsupervised classification approaches. Em: **Proceedings of the 38th annual Hawaii international conference on system sciences**, 2005.

CHAPELLE, O.; SCHOLKOPF, B.; ZIEN EDS., A. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. **IEEE Transactions on Neural Networks**, v. 20, n. 3, p. 542, 2009.

CHEN, C.; LIAW, A. **Using Random Forest to Learn Imbalanced Data**. 2004. Disponível em: <http://www.stat.berkeley.edu/tech-reports/666.pdf>. Acesso em: 31 ago. 2022.

CHEN, J.; GUO, X. J.; LI, H. Implementation and practice of an integrated process to recover copper from low grade ore at Zijinshan mine. **Hydrometallurgy**, v. 195, p. 105394, 2020.

CHOI, Y.; LEE, H. W. Trends in Mineral Resources Development Technology Using Artificial Intelligence. Em: **ITFIND**, 2020. p. 13–24.

CORTES, C.; VAPNIK, V.; SAITTA, L. Support-vector networks. **Machine Learning 1995 20:3**, v. 20, n. 3, p. 273–297, 1995.

COSTA, G.; ORTALE, R.; RITACCO, E. **Effective XML Classification Using Content and Structural Information via Rule Learning**. Em: 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011, [...]. IEEE Computer Society, 2011. p. 102–109.

COTTIS, R. A. Modelling corrosion in nuclear power plant systems. **Nuclear Corrosion Science and Engineering**, p. 438–448, 2012.

CS231N. **Convolutional Neural Networks for Visual Recognition**. Disponível em: <https://cs231n.github.io/neural-networks-1/>. Acesso em: 21 jul. 2022.

CSÁJI, B. C. Approximation with artificial neural networks. **Faculty of Sciences, Etvs Lornd University, Hungary**, v. 24, n. 48, p. 7, 2001.

DA SILVA, A. C. S.; DA COSTA, M. L. Genesis of the “soft” iron ore at S11D Deposit, in Carajás, Amazon Region, Brazil. **Brazilian Journal of Geology**, v. 50, n. 1, 2020.

DATAIKU DSS 11.0 DOCUMENTATION. **Reference API documentation of dataiku — Dataiku DSS 11.0 documentation**. Disponível em: <https://doc.dataiku.com/dss/latest/python-api/dataiku-reference.html>. Acesso em: 31 ago. 2022.

D′AZEREDO ORLANDO, M. T.; GALVÃO, E. S.; SANT′ANA CAVICHINI, A.; GABRIG TURBAY RANGEL, C. V.; PINHEIRO ORLANDO, C. G.; GRILO, C. F.; SOARES, J.; SANTOS OLIVEIRA, K. S.; SÁ, F.; JUNIOR, A. C.; BASTOS, A. C.; DA SILVA QUARESMA, V. Tracing iron ore tailings in the marine environment: An investigation of the Fundão dam failure. **Chemosphere**, v. 257, p. 127184, 2020.

DEO, A. J.; SAHOO, A.; BEHERA, S. K.; DAS, D. P. Machine Learning based Image Processing for Iron Ore Pellet Size Analysis. **2021 International Conference on Nascent Technologies in Engineering, ICNET 2021 - Proceedings**, 2021. Acesso em: 22 jul. 2022.

DOHERTY, S. **Control of pH in Chemical Processes Using Artificial Neural Networks**. 1999. Liverpool John Moores University, Liverpool, 1999. Disponível em: <https://www.researchgate.net/publication/265486784>. Acesso em: 15 jul. 2022.

DONG, J. J.; WANG, G.; GONG, Y. G.; XUE, Q. G.; WANG, J. S. Effect of high alumina iron ore of gibbsite type on sintering performance. **Ironmaking & Steelmaking**, v. 42, n. 1, p. 34–40, 2014.

EBNER, J. **Regression vs Classification, Explained - Sharp Sight**. Disponível em: <https://www.sharpsightlabs.com/blog/regression-vs-classification/>. Acesso em: 7 jul. 2022.

FALCONI, I. B. A.; BALTAZAR, M. dos P. G.; ESPINOSA, D. C. R.; TENÓRIO, J. A. S. Degradation of surfactant used in iron mining by oxidation technique: Fenton, photo-Fenton, and H2O2/UV - A comparative study. **The Canadian Journal of Chemical Engineering**, v. 98, n. 5, p. 1069–1083, 2020.

FAWAGREH, K.; GABER, M. M.; ELYAN, E. Random forests: from early developments to recent advancements. **Systems Science & Control Engineering: An Open Access Journal**, v. 2, n. 1, p. 602–609, 2014.

FLEMMER, R.; FLEMMER, C.; BRUNETTE, E. S.; FLEMMER, R. C.; FLEMMER, C. L. A review of artificial intelligence. 2009. In: **2009 4th International Conference on Autonomous Robots and Agents**. Ieee, 2009. p. 385-392.

FLORES, V.; KEITH, B.; LEIVA, C. Using Artificial Intelligence Techniques to Improve the Prediction of Copper Recovery by Leaching. **Journal of Sensors**, 2020.

FOSTER, H.; GOLIYA, K. **Analysis: Iron ore disruptions in southern Brazil may support recovery in spot prices | S&P Global Commodity Insights**. Petróleo Metais Produtos Petroquímicos, 10 jan. 2022. Disponível em: <https://www.spglobal.com/commodityinsights/pt/market-insights/latest-news/metals/011022-analysis-iron-ore-disruptions-in-southern-brazil-may-support-recovery-in-spot-prices>. Acesso em: 22 jul. 2022.

FRANCOIS-LAVET, V.; HENDERSON, P.; ISLAM, R.; BELLEMARE, M. G.; PINEAU, J. An Introduction to Deep Reinforcement Learning. **Foundations and Trends in Machine Learning**, v. 11, n. 3–4, p. 219–354, 2018.

GHORBANZADEH, O.; BLASCHKE, T. Optimizing sample patches selection of CNN to improve the MIOU on landslide detection. **GISTAM 2019 - Proceedings of the 5th International Conference on Geographical Information Systems Theory, Applications and Management**, p. 33–40, 2019.

GLOBAL DATA. **Leading AI Companies in Mining & Mineral Exploration - Mining Technology**. Disponível em: <https://www.mining-technology.com/buyers-guide/leading-ai-companies-mining/>. Acesso em: 16 jul. 2022.

GOLDSTEIN, E. B.; COCO, G. Machine learning components in deterministic models: Hybrid synergy in the age of data. **Frontiers in Environmental Science**, v. 3, n. APR, p. 33, 2015.

GOMES, R. B.; TOMI, G. D.; ASSIS, P. S. Impact of quality of iron ore lumps on sustainability of mining operations in the Quadrilatero Ferrifero Area. **Minerals Engineering**, v. 70, p. 201–206, 2015.

GOOD, I. J. The Philosophy of Exploratory Data Analysis. **Philosophy of Science**, v. 50, n. 2, p. 283–295, 1983.

GOSWAMI, R. Comparison of state-of-the-art machine learning based data driven and model updating methods against shallow and deep convolutional neural networks methods of structural damage. **International Research Journal of Engineering and Technology (IRJET),** v. 7, n. 12, p. 1494–1528, 2021.

GRANEK, J.; HABER, E. Data mining for real mining: A robust algorithm for prospectivity mapping with uncertainties. Em: **Proceedings of the 2015 SIAM international conference on data mining**. Society for Industrial and Applied Mathematics, 2015. p. 145-153**.**

GREENGARD, P. The Neurobiology of Slow Synaptic Transmission. **Science**, v. 294, n. 5544, p. 1024–1030, 2001.

HAYKIN, S. **Redes neurais: princípios e prática**. [s.l.] Bookman Editora, 2001.

HE, Y.; ZHOU, Y.; WEN, T.; ZHANG, S.; HUANG, F.; ZOU, X.; MA, X.; ZHU, Y. A review of machine learning in geochemistry and cosmochemistry: Method improvements and applications. **Applied Geochemistry**, v. 140, p. 105273, 2022.

HILL, E. J.; FABRIS, A.; UVAROVA, Y.; TIDDY, C. Improving geological logging of drill holes using geochemical data and data analytics for mineral exploration in the Gawler Ranges, South Australia. **Australian Journal of Earth Sciences**, p. 1-27, 2021.

HOOD, S. B.; CRACKNELL, M. J.; GAZLEY, M. F. Linking protolith rocks to altered equivalents by combining unsupervised and supervised machine learning. **Journal of Geochemical Exploration**, v. 186, p. 270–280, 2018.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. **Neural Networks**, v. 2, n. 5, p. 359–366, 1989.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science and Engineering**, v. 9, n. 3, p. 90–95, 2007.

IBRAM. **PANORAMA DA MINERAÇÃO EM MINAS GERAIS**. Brasília, 2015. Acesso em: 22 jul. 2022.

IBRAM. **Anuário divulga balanço do setor mineral em 2018 - IBRAM**. Disponível em: <https://ibram.org.br/noticia/anuario-divulga-balanco-do-setor-mineral-em-2018/>. Acesso em: 22 jul. 2022.

IPEA. **Diagnóstico dos Resíduos Sólidos da Atividade de Mineração de Substâncias Não Energéticas**. 2012. Disponível em: <http://www.ipea.gov.br>. Acesso em: 24 jul. 2022.

JORDAN, C.; ZHANG, C.; HIGGINS, A. Using GIS and statistics to study influences of geology on probability features of surface soil geochemistry in Northern Ireland. **Journal of Geochemical Exploration**, v. 93, n. 3, p. 135–152, 2007.

JUNG, D.; CHOI, Y. Systematic Review of Machine Learning Applications in Mining: Exploration, Exploitation, and Reclamation. **Minerals 2021, Vol. 11, Page 148**, v. 11, n. 2, p. 148, 2021.

KAPOOR, S.; NARAYANAN, A. Leakage and the Reproducibility Crisis in ML-based Science. **arXiv preprint arXiv:2207.07048**, 2022.

KEYKHAY-HOSSEINPOOR, M.; KOHSARY, A. H.; HOSSEIN-MORSHEDY, A.; PORWAL, A. A machine learning-based approach to exploration targeting of porphyry Cu-Au deposits in the Dehsalm district, eastern Iran. **Ore Geology Reviews**, v. 116, p. 103234, 2020.

KOST, S.; RHEINBACH, O.; SCHAEBEN, H. Using logistic regression model selection towards interpretable machine learning in mineral prospectivity modeling. **Geochemistry**, v. 81, n. 4, p. 125826, 2021.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Machine learning: a review of classification and combining techniques. **Artificial Intelligence Review 2007 26:3**, v. 26, n. 3, p. 159–190, 2007.

KRATSIOS, A.; BILOKOPYTOV, I. Non-euclidean universal approximation. **Advances in Neural Information Processing Systems**, v. 33, p. 10635–10646, 2020.

LARY, D. J. Artificial Intelligence in Geoscience and Remote Sensing. **Geoscience and Remote Sensing New Achievements**, p. 105, 2010.

LARY, D. J.; ALAVI, A. H.; GANDOMI, A. H.; WALKER, A. L. Machine learning in geosciences and remote sensing. **Geoscience Frontiers**, v. 7, n. 1, p. 3–10, 2016.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature 2015 521:7553**, v. 521, n. 7553, p. 436–444, 2015.

LEE, T.-H.; ULLAH, A.; WANG, R. Bootstrap aggregating and random forest. *Em*: **Macroeconomic forecasting in the era of big data**. Springer, Cham, 2020. p. 389-429.

LI, S.; CHEN, J.; LIU, C.; WANG, Y. Mineral Prospectivity Prediction via Convolutional Neural Networks Based on Geological Big Data. **Journal of Earth Science,** v. 32, n. 2, p. 327–347, 2021.

LI, S.; CHEN, J.; XIANG, J. Applications of deep convolutional neural networks in prospecting prediction based on two-dimensional geological big data. **Neural Computing and Applications,** v. 32, n. 7, p. 2037–2053, 2019.

LIMA, R. M. F.; ABREU, F. D. P. V. F. Characterization and concentration by selective flocculation/magnetic separation of iron ore slimes from a dam of Quadrilátero Ferrífero - Brazil. **Journal of Materials Research and Technology**, v. 9, n. 2, p. 2021–2027, 2020.

MACHADO, G.; MENDOZA, M. R.; CORBELLINI, L. G. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. **Veterinary Research**, v. 46, n. 1, 2015.

MARTINS, P. F. F.; MORAIS, C. A.; LAMEIRAS, F. S.; ALBUQUERQUE, R. O.; MARTINS, P. F. F.; MORAIS, C. A.; LAMEIRAS, F. S.; ALBUQUERQUE, R. O. Silica and Iron Recovery from a Residue of Iron Ore Flotation. **Journal of Minerals and Materials Characterization and Engineering**, v. 5, n. 4, p. 153–160, 2017.

MATIOLO, E.; COUTO, H. J. B.; LIMA, N.; SILVA, K.; DE FREITAS, A. S. Improving recovery of iron using column flotation of iron ore slimes. **Minerals Engineering**, v. 158, p. 106608, 2020.

MERDITH, A. S.; LANDGREBE, T. C. W.; MULLER, R. D.; MÜLLER, R. D. Building a machine learning classifier for iron ore prospectivity in the Yilgarn Craton. **ASEG Extended Abstracts**, v. 2015, n. 1, p. 1-4, 2015.

MEREMBAYEV, T.; YUNUSSOV, R.; YEDILKHAN, A. Machine learning algorithms for classification geology data from well logging. **14th International Conference on Electronics Computer and Computation, ICECCO 2018**, 2019.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine Learning: an Artificial Intelligence Approach**. Germany: Springer Berlin Heidelberg, 1983. 572 p.

MILITKÝ, J. Fundamentals of soft models in textiles. **Soft Computing in Textile Engineering**, p. 45–102, 1 jan. 2011.

MIN, D. H.; YOON, H. K. Suggestion for a new deterministic model coupled with machine learning techniques for landslide susceptibility mapping. **Scientific Reports,** v. 11, n. 1, p. 1–24, 2021.

MOLUGARAM, K.; RAO, G. S.; MOLUGARAM, K.; RAO, G. S. MOLUGARAM, K. et al. Chapter 5 - Curve Fitting. **Statistical Techniques for Transportation Engineering; Butterworth-Heinemann: Woburn, MA, USA**, p. 281-292, 2017.

NAKHAEI, F.; IRANNAJAD, M. Reagents types in flotation of iron oxide minerals: A review. **https://doi.org/10.1080/08827508.2017.1391245**, v. 39, n. 2, p. 89–124, 2017.

NASCIMENTO, M. G. L. **A Dinâmica do Mercado do Minério de Ferro no Comércio Internacional**. 2021. Disponível em: < https://repositorio.pucgoias.edu.br/jspui/handle/123456789/2630>. Acesso em: 22 jul. 2022.

NIELSEN, M. A. **Neural Networks and Deep Learning**. San Francisco, CA, USA: Determination Press, 2015.

NOBLE, W. S. What is a support vector machine? **Nature biotechnology**, v. 24, n. 12, p. 1565-1567, 2006.

ODELL, S. D.; BEBBINGTON, A.; FREY, K. E. Mining and climate change: A review and framework for analysis. **The Extractive Industries and Society**, v. 5, n. 1, p. 201–214, 2018.

OKADA, N.; MAEKAWA, Y.; OWADA, N.; HAGA, K.; SHIBAYAMA, A.; KAWAMURA, Y. Automated Identification of Mineral Types and Grain Size Using Hyperspectral Imaging and Deep Learning for Mineral Processing. **Minerals 2020, Vol. 10, Page 809**, v. 10, n. 9, p. 809, 2020.

OLIVEIRA, S. S. 2022. **DOI:10.5281/zenodo.6968443**.

PARMAR, A.; KATARIYA, R.; PATEL, V. A Review on Random Forest: An Ensemble Classifier. **Lecture Notes on Data Engineering and Communications Technologies**, v. 26, p. 758–763, 2019

POLIKAR, R. Ensemble Learning. **Ensemble Machine Learning**, p. 1–34, 2012.

PROBST, P.; BOULESTEIX, A. L. To Tune or Not to Tune the Number of Trees in Random Forest. **Journal of Machine Learning Research**, v. 18, p. 1–18, 2018.

QISHUAI, Y.; JIN, Y.; BO, Z.; MENGLEI, J.; XIAOLIANG, C.; CHAO, F.; LI, Y.; LEI, L.; YATAO, L.; ZHENGLI, L. Improve the Drilling Operations Efficiency by the Big Data Mining of Real-Time Logging. **Proceedings of the SPE/IADC Middle East Drilling Technology Conference and Exhibition**, 2018.

RAGHUKUMAR, C.; KUMAR TRIPATHY, S.; MOHANAN, S. Beneficiation of Indian High Alumina Iron Ore Fines – a Case Study. **International Journal of Mining Engineering and Mineral Processing**, v. 1, n. 2, p. 94–100, 2012.

RAHMAN, A.; SHAHRIAR, M. S.; TIMMS, G.; LINDLEY, C.; DAVIE, A. B.; BIGGINS, D.; HELLICAR, A.; SENNERSTEN, C.; SMITH, G.; COOMBE, M. A machine learning approach to find association between imaging features and XRF signatures of rocks in underground mines. Em: **2015 IEEE SENSORS**. IEEE, 2015. p. 1-4.

RAHMAN, A.; TIMMS, G.; SHAHRIAR, M. S.; SENNERSTEN, C.; DAVIE, A.; LINDLEY, C. A.; HELLICAR, A. D.; SMITH, G.; BIGGINS, D.; COOMBE, M. Association Between Imaging and XRF Sensing: A Machine Learning Approach to Discover Mineralogy in Abandoned Mine Voids. **IEEE Sensors Journal**, v. 16, n. 11, p. 4555–4565, 2016.

REBACK, J.; JBROCKMENDEL; MCKINNEY, W.; BOSSCHE, J. van den; ROESCHKE, M.; AUGSPURGER, T.; HAWKINS, S.; CLOUD, P.; GFYOUNG; SINHRKS; HOEFLER, P.; KLEIN, A.; PETERSEN, T.; TRATNER, J.; SHE, C.; AYD, W.; NAVEH, S.; DARBYSHIRE, J.; SHADRACH, R.; GARCIA, M.; SCHENDEL, J.; HAYDEN, A.; SAXTON, D.; GORELLI, M. E.; LI, F.; WÖRTWEIN, T.; ZEITLIN, M.; JANCAUSKAS, V.; MCMASTER, A.; LI, T. pandas-dev/pandas: Pandas 1.0. 5. **Zenodo**, 2020.

RODRIGUEZ-GALIANO, V. F.; CHICA-OLMO, M.; CHICA-RIVAS, M. Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar

area, Southern Spain. **International Journal of Geographical Information Science**, v. 28, n. 7, p. 1336-1354, 2014.

RODRIGUEZ-GALIANO, V. F.; GHIMIRE, B.; ROGAN, J.; CHICA-OLMO, M.; RIGOL-SANCHEZ, J. P. An assessment of the effectiveness of a random forest classifier for land-cover classification. **ISPRS journal of photogrammetry and remote sensing**, v. 67, p. 93-104, 2012.

RODRIGUEZ-GALIANO, V.; SANCHEZ-CASTILLO, M.; CHICA-OLMO, M.; CHICA-RIVAS, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. **Ore Geology Reviews**, v. 71, p. 804–818, 1 dez. 2015.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386–408, 1958.

ROY, S. K.; NAYAK, D.; RATH, S. S. A review on the enrichment of iron values of low-grade Iron ore resources using reduction roasting-magnetic separation. **Powder Technology**, v. 367, p. 796–808, 2020.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Internal Representations by Error Propagation. Em: RUMELHART, D. E.; MCCLELLAND. **Learning internal representations by error propagation**. Parallel distributed processing: MIT Press, 1986. p. 318–362.

SAH, S. **Machine Learning: A Review of Learning Types**. 2020. Disponível em < https://www.preprints.org/manuscript/202007.0230/v1>. Acesso em 15 jul. 2022.

SAHOO, K.; SAMAL, A. K.; PRAMANIK, J.; PANI, S. K. Exploratory data analysis using Python. **International Journal of Innovative Technology and Exploring Engineering (IJITEE)**, v. 8, n. 12, p. 2019, 2019.

SANTOS, B. C. **Avaliação da Eficiência da Deslamagem de Minério de Ferro Via Decantação e Hidrociclonagem**. 2018. Disponível em: <https://www.eng-minas.araxa.cefetmg.br/wp-content/uploads/sites/170/2018/05/Bianca-de-Castro-Santos.pdf>. Acesso em: 22 jul. 2022.

SAROUFIM, C. E. **Internet of Things and Anomaly Detection for the Iron Ore Mining Industry**. 2016. Massachusetts Institute of Technology, Cambridge, 2016.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural Networks**, v. 61, p. 85–117, 1 jan. 2015.

SCHNITZLER, N.; ROSS, P. S.; GLOAGUEN, E. Using machine learning to estimate a key missing geochemical variable in mining exploration: Application of the Random Forest algorithm to multi-sensor core logging data. **Journal of Geochemical Exploration**, v. 205, p. 106344, 2019.

SCHÖLKOPF, B. SVMs - A practical consequence of learning theory. **IEEE Intelligent Systems and Their Applications**, v. 13, n. 4, p. 18–21, 1998.

SCHÖNIG, J.; VON EYNATTEN, H.; TOLOSANA-DELGADO, R.; MEINHOLD, G. Garnet major-element composition as an indicator of host-rock type: a machine learning approach using the random forest classifier. **Contributions to Mineralogy and Petrology**, v. 176, n. 12, p. 1–21, 2021.

SCHUNNESSON, H. **Drill process monitoring in percussive drilling: A multivariate approach to data analysis**. 1990. Lulea University of Technology, Lulea, Sweden, 1990. Disponível em: <http://ltu.diva-portal.org/smash/get/diva2:990596/FULLTEXT01.pdf>. Acesso em: 31 ago. 2022.

SHAHABI, H.; JARIHANI, B.; PIRALILOU, S. T.; CHITTLEBOROUGH, D.; AVAND, M.; GHORBANZADEH, O. A Semi-Automated Object-Based Gully Networks Detection Using Different Machine Learning Models: A Case Study of Bowen Catchment, Queensland, Australia. **Sensors (Basel, Switzerland)**, v. 19, n. 22, 2019.

SHENG, L.; ZHANG, T.; NIU, G.; WANG, K.; TANG, H.; DUAN, Y.; LI, H. Classification of iron ores by laser-induced breakdown spectroscopy (LIBS) combined with random forest (RF). **Journal of Analytical Atomic Spectrometry**, v. 30, n. 2, p. 453–458, 28 jan. 2015.

SINAICE, B. B.; KAWAMURA, Y.; SHIBUYA, T.; SASAKI, J.; YOSHIMOTO, H.; ITO, Y.; UTSUKI, S. Development of a differentiation and identification system for igneous rocks using hyper-spectral images and a convolutional neural network (CNN) system. Em: **Proceedings of the MMIJ**. Sapporo, Japan: 2017.

STONE, M. Cross-Validatory Choice and Assessment of Statistical Predictions. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 36, n. 2, p. 111–147, 1974.

SUN, T.; CHEN, F.; ZHONG, L.; LIU, W.; WANG, Y. GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China. **Ore Geology Reviews**, v. 109, p. 26–49, 2019.

TESSEMA, A. Mineral Systems Analysis and Artificial Neural Network Modeling of Chromite Prospectivity in the Western Limb of the Bushveld Complex, South Africa. **Natural Resources Research,** v. 26, n. 4, p. 465–488, 2017.

THOMPSON, F.; DE OLIVEIRA, B. C.; CORDEIRO, M. C.; MASI, B. P.; RANGEL, T. P.; PAZ, P.; FREITAS, T.; LOPES, G.; SILVA, B. S.; S. CABRAL, A.; SOARES, M.; LACERDA, D.; DOS SANTOS VERGILIO, C.; LOPES-FERREIRA, M.; LIMA, C.; THOMPSON, C.; DE REZENDE, C. E. Severe impacts of the Brumadinho dam failure (Minas Gerais, Brazil) on the water quality of the Paraopeba River. **The Science of the total environment**, v. 705, 2020.

TOHRY, A.; JAFARI, M.; FARAHANI, M.; MANTHOURI, M.; CHELGANI, S. C. Variable importance assessments of an innovative industrial-scale magnetic separator for processing of iron ore tailings. **Mineral Processing and Extractive Metallurgy: Transactions of the Institute of Mining and Metallurgy**, v. 131, n. 2, p. 122–129, 2022.

VANSCHOREN, J.; BLOCKEEL, H. Experiment databases. **Inductive Databases and Constraint-Based Data Mining**, p. 335–361, 2010.

VERONESI, F.; SCHILLACI, C. Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. **Ecological Indicators**, v. 101, p. 1032–1044, 2019.

WANG, J.; ZHOU, Y.; XIAO, F. Identification of multi-element geochemical anomalies using unsupervised machine learning algorithms: A case study from Ag–Pb–Zn deposits in north-western Zhejiang, China. **Applied Geochemistry**, v. 120, 2020.

WILLMOTT, C. J. Some comments on the evaluation of model performance. **Bulletin of the American Meteorological Society**, v. 63, n. 11, p. 1309–1313, 1982.

XIONG, Y.; ZUO, R.; CARRANZA, E. J. M. Mapping mineral prospectivity through big data analytics and a deep learning algorithm. **Ore Geology Reviews**, v. 102, p. 811–817, 2018.

XU, Y.; GOODACRE, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. **Journal of Analysis and Testing**, v. 2, n. 3, p. 249–262, 2018.

YANG, C.; CUI, C.; QIN, J.; CUI, X. Characteristics of the fired bricks with low-silicon iron tailings. **Construction and Building Materials**, v. 70, p. 36–42, 2014.

ZHANG, C.; QIN, Y.; ZHU, X.; ZHANG, J.; ZHANG, S. **Clustering-based missing value imputation for data preprocessing**. Em: 4th IEEE International Conference on Industrial Informatics. IEEE, 2006. p. 1081-1086.

ZARE, M.; HUOVA, M.; VISA, A.; LAUNIS, S. Real-time online drilling vibration analysis using data mining. **ACM International Conference Proceeding Series**, p. 175–180, 2019.

ZHANG, N.; ZHOU, K.; LI, D. Back-propagation neural network and support vector machines for gold mineral prospectivity mapping in the Hatu region, Xinjiang, China. **Earth Science Informatics 2018 11:4**, v. 11, n. 4, p. 553–566, 2018.

ZHANG, S.; XIAO, K.; CARRANZA, E. J. M.; YANG, F.; ZHAO, Z. Integration of auto-encoder network with density-based spatial clustering for geochemical anomaly detection for mineral exploration. **Computers and Geosciences**, v. 130, p. 43–56, 2019.

ZHANG, T.; XIA, D.; TANG, H.; YANG, X.; LI, H. Classification of steel samples by laser-induced breakdown spectroscopy and random forest. **Chemometrics and Intelligent Laboratory Systems**, v. 157, p. 196–201, 2016.

ZHANG, Z.; CHENG, Q.; YANG, J.; WU, G.; GE, Y. Machine learning for mineral prospectivity: A case study of iron-polymetallic mineral prospectivity in southwestern Fujian. **Earth Science Frontiers**, v. 28, n. 3, p. 221, 2021.

ZHAO, J.; CHEN, S. Identification of the Ore-Forming Anomaly Component by MSVD Combined with PCA from Element Concentrations in Fracture Zones of the Laochang Ore Field, Gejiu, SW China. **Journal of Earth Science 2021 32:2**, v. 32, n. 2, p. 427–438, 2021.

ZUO, R. Machine Learning of Mineralization-Related Geochemical Anomalies: A Review of Potential Methods. **Natural Resources Research 2017 26:4**, v. 26, n. 4, p. 457–464, 2017.

ZUO, R. Mineral Exploration Using Subtle or Negative Geochemical Anomalies. **Journal of Earth Science 2021 32:2**, v. 32, n. 2, p. 439–454, 2020.

ZUO, R.; CARRANZA, E. J. M. Support vector machine: A tool for mapping mineral prospectivity. **Computers & Geosciences**, v. 37, n. 12, p. 1967–1975, 2011.

ZUO, R.; ZUO, R. Mineral Exploration Using Subtle or Negative Geochemical Anomalies. **Journal of Earth Science, 2021, Vol. 32, Issue 2, Pages: 439-454**, v. 32, n. 2, p. 439–454, 2021.