INSTITUTO TECNOLÓGICO VALE

Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração (PROFICAM)
Escola de Minas, Universidade Federal de Ouro Preto (UFOP)
Associação Instituto Tecnológico Vale (ITV)

Dissertação

ESTIMANDO TEORES DE FERRO EM MINÉRIOS: UMA INVESTIGAÇÃO COM MÉTODOS DE APRENDIZADO DE MÁQUINA E IMAGENS HIPERESPECTRAIS

Arthur Oliveira Viana

Ouro Preto
Minas Gerais, Brasil
2020

Arthur Oliveira Viana

ESTIMANDO TEORES DE FERRO EM MINÉRIOS: UMA INVESTIGAÇÃO COM MÉTODOS DE APRENDIZADO DE MÁQUINA E IMAGENS HIPERESPECTRAIS

Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração da Universidade Federal de Ouro Preto e do Instituto Tecnológico Vale, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Controle e Automação.

Orientador: Prof. Gustavo Pessin, D.Sc.
Coorientadora: Profª. Rosa Elvira Correa Pabón, D.Sc.

Ouro Preto
2020

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

V614e Viana, Arthur Oliveira .
Estimando teores de ferro em minérios [manuscrito]: uma investigação com métodos de aprendizado de máquina e imagens hiperespectrais. / Arthur Oliveira Viana. - 2020.
66 f.

Orientador: Prof. Dr. Gustavo Pessin.
Coorientadora: Profa. Dra. Rosa Elvira Correa Pabón.
Dissertação (Mestrado Profissional). Universidade Federal de Ouro Preto. Programa de Mestrado Profissional em Instrumentação, Controle e Automação de Processos de Mineração. Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração.
Área de Concentração: Engenharia de Controle e Automação de Processos Minerais.

1. Aprendizado do computador . 2. Minas e recursos minerais - Caracterização mineral. 3. Imagem hiperespectral. I. Pabón, Rosa Elvira Correa . II. Pessin, Gustavo. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 681.5:622.2

Bibliotecário(a) Responsável: Maristela Sanches Lima Mesquita - CRB:1716

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
PROGR. POS GRAD. PROF. INST. CONT. E AUT. PROCESSOS DE MIN.

FOLHA DE APROVAÇÃO

Arthur Oliveira Viana

Estimando Teores de Ferro em Minérios: Uma Investigação com Métodos de Aprendizado de Máquina e Imagens Hiperespectrais

Membros da banca

Vidal Félix Navarro Torres - Ph.D. - Instituto Tecnológico Vale Mineração

Luiz Gonzaga da Silveira Junior - Dr. - Universidade do Vale do Rio dos Sinos

Paulo Antônio de Souza Júnior - Dr. rer. nat. - Griffith University

Versão final
Aprovada em 07 de Julho de 2020

De acordo

Gustavo Pessin - Dr. - Instituto Tecnológico Vale Mineração (Orientador)

Rosa Elvira Correa Pabón - Dra. - Instituto Tecnológico Vale Mineração (Coorientadora)

Documento assinado eletronicamente por **Agnaldo Jose da Rocha Reis**, **COORDENADOR DO CURSO DE POS-GRADUACAO EM INSTRUMENTACAO, CONTROLE E AUTOMACAO DE PROC DE MINERACAO**, em 13/07/2020, às 13:00, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0066717** e o código CRC **69E314C0**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: - www.ufop.br

À minha família e amigos, pelo incentivo e compreensão nas horas de ausência.

Agradecimentos

Agradeço a minha família e amigos por sempre estarem presentes e darem o apoio e suporte necessários a conclusão desta etapa. Agradeço também ao meu orientador Gustavo Pessin e coorientadora Rosa Elvira Correa Pabón pela paciência, ensinamentos e por confiarem no meu potencial. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES), Código de Financiamento 001; do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG); e da Vale SA.

"O futuro pertence àqueles que se preparam hoje para ele"
(Malcolm X).

Resumo

Resumo da Dissertação apresentada ao Programa de Pós Graduação em Instrumentação, Controle e Automação de Processos de Mineração como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

ESTIMANDO TEORES DE FERRO EM MINÉRIOS: UMA INVESTIGAÇÃO COM MÉTODOS DE APRENDIZADO DE MÁQUINA E IMAGENS HIPERESPECTRAIS

Arthur Oliveira Viana

Julho/2020

Orientadores: Gustavo Pessin
Rosa Elvira Correa Pabón

Processos de beneficiamento mineral como prospecção, pesquisa, lavra e beneficiamento mineral podem ganhar com processos mais ágeis de caracterização dos minérios. A caracterização, feita por métodos tradicionais em laboratório, é muito precisa, mas em geral apresenta deficiência de tempo. A análise de imagens hiperespectrais pode trazer resultados mais rápidos do que a análise tradicional em laboratório, entretanto, a precisão da caracterização ainda é um desafio a ser investigado. Estas dificuldades têm relação com fatores ambientais como iluminação e umidade, fatores amostrais como tamanho e homogeneidade dos grãos, e fatores de modelagem, como escolha de bandas espectrais, resolução de imagens e tipos de modelos para caracterização. Considerando os desafios citados, esta pesquisa objetivou responder questões relacionadas aos fatores de modelagem e, portanto, investigamos métodos de aprendizado de máquina para estimar o teor de ferro em amostras de minérios de ferro com base em comprimentos de onda de imagens hiperespectrais na região do Visible and near infrared (VNIR) entre 400 e 1000 *nm*; realizamos uma seleção dos atributos mais relevantes para o modelo e validamos os resultados com o uso de métricas de avaliação estatísticas. O desempenho dos modelos manifestou resultados constantes, que apresentam baixa variância e dispersão e com precisão de estimação dos teores de ferro acima de 90% utilizando Random Forests (RF) e Multilayer Perceptrons (MLP).

Palavras-chave: Aprendizado de máquina, Caracterização mineral, Imagem hiperespectral.

Macrotema: Mina; **Linha de Pesquisa:** Tecnologias da Informação, Comunicação e Automação Industrial; **Tema:** Redução de incerteza no planejamento da mina.

Abstract

Abstract of Dissertation presented to the Graduate Program on Instrumentation, Control and Automation of Mining Process as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

ESTIMATING IRON CONTENT IN ORES: AN INVESTIGATION WITH MACHINE LEARNING METHODS AND HYPERSPECTRAL IMAGES

Arthur Oliveira Viana

July/2020

Advisors: Gustavo Pessin
Rosa Elvira Correa Pabón

Mineral processing processes such as mineral exploration activities can gain from more agile processes for characterizing ores. The characterization, carried out by traditional methods in the laboratory, is very accurate, but in general it is time deficient. The analysis of hyperspectral images can bring faster results than traditional laboratory analysis, however, the accuracy of the characterization is still a challenge to be investigated. These difficulties are related to environmental factors such as lighting and humidity, sample factors such as grain size and homogeneity, and modeling factors, such as choice of spectral bands, resolution of images and types of models for characterization. Considering the challenges mentioned, this research aimed to answer questions related to the modeling factors and therefore, we investigated machine learning methods to estimate the iron content in iron ore samples based on wavelengths of hyperspectral images in the Visible and near infrared region (VNIR) between 400 and 1000 *nm*; most relevant attributes for the model and we validate the results with the use of statistical evaluation metrics. The performance of the models showed constant results, which present low variance and dispersion and with iron dosage estimation accuracy above 90% using Random Forests (RF) and Multilayer Perceptrons (MLP).

Keywords: Machine learning, Mineral characterization, Hyperspectral image.

Macrotheme: Mine; **Research Line:** Information Technologies, Communication and Industrial Automation; **Theme:** Uncertainty reduction in mine planning.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

AAP Algoritmos de Aprendizado de Máquina

BPNN Modelos de Rede Neural de Retropropagação

ELM Máquina de Aprendizado Extremo

IA Inteligência Artificial

LS-SVM Modelos de Máquina de Vetores de Suporte de Mínimos Quadrados

LWIR Long-wavelength Infrared

MAE Erro Médio Absoluto

MEV Microscópio Eletrônico de Varredura

MLP Multilayer Perceptron

MSE Erro Médio Quadrático

MWIR Mid-wavelength Infrared

NIR Near Infrared

R^2 Coeficiente de Determinção

RF Random Forest

RNA Redes Neurais Artificiais

SVM Support Vector Machines

SWIR Short-wavelength Infrared

VNIR Visible and Near Infrared

Lista de Símbolos

Å angstron

θ ângulo de Bragg

λ comprimento de onda dos raios no retículo refratado

d distância dos átomos

g grama

n ordem de difração

sen seno

1. Introdução

Este capítulo busca introduzir o tema da pesquisa de maneira sucinta apresentando conceitos gerais, assim como expor a contextualização, justificava do projeto e objetivos gerais e específicos. Ademais, é disposta a estrutura do texto em seus capítulos com uma breve descrição dos conteúdos.

1.1. Contextualização e justificativa

A caracterização de minérios (mineralógica, espectral ou química) é realizada com a finalidade de se obter dados que serão úteis no desenvolvimento de diferentes fases operacionais de minério. Tal procedimento não pode deixar de apresentar uma quantificação de seus constituintes representados pelos minerais de interesse econômico e pelos elementos paragêneses. Segundo Neumann *et al.* (2010), o modo de se caracterizar uma amostra de minério varia muito com a própria mineralogia e suas propriedades inerentes, bem como os objetivos, a abrangência da caracterização, as possíveis rotas operacionais, a disponibilidade de tempo, a capacidade analítica e os recursos financeiros. A quantidade de minério utilizada na análise deve ser cuidadosamente manipulada para que sua representatividade possa prevalecer em todas as etapas da caracterização. Este cuidado começa com a amostragem em campo e deve ser mantido inclusive em testes de escala laboratorial. Uma amostragem cuidadosa envolve critérios de homogeneidade, conhecimento aproximado do peso do mineral de interesse econômico e seu teor de ferro (GY, 2012).

A identificação e quantificação das fases mineralógicas constituintes dos minérios de ferro é um tema de grande importância para a área de mineralogia, desta forma, faz-se necessário o estudo de diversas técnicas que possam agregar maior eficiência à prospecção mineral, como por exemplo, o sensoriamento remoto hiperespectral. Existem vários artigos de revisão sobre o tema de sensoriamento remoto e imageamento hiperespectral, tanto para exploração mineral quanto para aplicações visando monitoramento ambiental. Em trabalhos recentes, Scafutto *et al.* (2017) indicam o crescente desenvolvimento de novas tecnologias, bem como o uso de dados de sensoriamento remoto hiperespectral como uma ferramenta complementar para exploração de petróleo e monitoramento ambiental na indústria petrolífera. Scafutto *et al.* (2017) também citam que a combinação da informação espectral com técnicas estatísticas oferece o potencial para aprimorar programas de exploração focados na descoberta de novas jazidas minerais.

Segundo Adão *et al.* (2017), o imageamento hiperespectral já não é mais uma técnica emergente, diversas pesquisas de impacto, a serem citadas no decorrer do texto, foram publicadas nas últimas décadas. Para cada pixel em uma imagem, uma câmera hiperespectral captura a energia de interação entre a radiação eletromagnética e o alvo de interesse. Em geral, cada pixel da imagem contém um espectro contínuo (a continuidade está relacionada com o número

de bandas da câmera) e pode ser usado para caracterizar os objetos com precisão. Dong *et al.* (2019) ressaltam que o desenvolvimento da tecnologia de espectroscopia de imagem tem feito um rápido progresso, focando principalmente na análise de características espectrais de minerais, pré-processamento de dados hiperespectrais e extração de informações. Entretanto, ainda é um desafio alcançar resultados com alta precisão quando considerados aspectos ambientais como variação da iluminação, umidade, etc.

Chen *et al.* (2008) ressaltam que o uso da inteligência artificial tem aumentado nos últimos anos nas áreas de modelagem ambiental, pelo fato da obtenção de resultados cada vez mais satisfatórios. Algoritmos de aprendizado de máquina como Redes Neurais Artificiais (RNA) e Random Forest (RF) são poderosos métodos baseados em dados que podem ser utilizados no mapeamento da prospectividade mineral. Rodriguez-Galiano *et al.* (2015) demonstram que os algoritmos de aprendizado de máquina são mais precisos do que as técnicas estatísticas, como a análise discreta ou a regressão logística, especialmente quando o problema é complexo. Esses métodos podem lidar com um grande número de características que são importantes em estudos de prospecção mineral. Alguns exemplos de aplicações desses métodos são: o trabalho de Leite e de Souza Filho (2009), que revisa o uso de RNAs aplicadas no mapeamento de potenciais minerais para mineralizações de cobre e ouro; e o trabalho de Abedi *et al.* (2012) que discute o uso de Support Vector Machines (SVM) para multiclassificação de áreas de prospecção mineral. Para otimizar o processo de análise de dados hiperespectrais, Hu *et al.* (2019) introduziram o conceito de seleção de bandas espectrais, que é um método efetivo para remover bandas redundantes e preservar a significância física do objeto estudado ao mesmo tempo. Numerosas bandas redundantes aumentam a complexidade computacional e o custo de armazenamento. Portanto, é essencial reduzir a dimensão dos dados de imagens hiperespectrais para evitar tais problemas.

Este trabalho se insere neste contexto, de buscar alternativas para a caracterização mineral por meio da aplicação de inteligência artificial e imageamento hiperespectral. Esta pesquisa foi realizada no intuito de trazer resultados quantitativos e qualitativos sobre as estimativas de teor de ferro em amostras e apresentar que o uso de técnicas de aprendizado de máquina é uma abordagem adequada ao problema.

1.2. Objetivos

O objetivo deste trabalho é estimar o teor de ferro em amostras de minérios com base em imagens hiperespectrais e inteligência artificial. Para tal, métodos de aprendizado de máquina são propostos, desenvolvidos e avaliados. As principais perguntas a serem respondidas neste trabalho são: (1) Qual o método mais preciso para a estimativa de teor de ferro usando imagem hiperespectral: Random Forest ou Multilayer Perceptron? (2) Como o número de amostras afeta o aprendizado dos métodos? e (3) Quais bandas da imagem hiperespectral são mais relevantes para a estimativa de teor de ferro? Para atingir os objetivos, um levantamento sobre conceitos

de mineralogia e um estudo sobre análise de dados químicos foi realizado (apresentado no Capítulo 2), além disso, foi realizada a coleta de imagens hiperespectrais em laboratório.

1.2.1. Objetivos específicos

- Propor, desenvolver e avaliar variações de modelos baseados em aprendizado de máquina para estimar teor de ferro em amostras de minérios considerando imagens hiperespectrais.
- Investigar o efeito do número de amostras em relação ao aprendizado dos métodos.
- Propor, desenvolver e avaliar e variações de métodos de seleção de atributos a fim de identificar as bandas mais relevantes.

1.3. Estrutura da dissertação

A presente dissertação está disposta em seis capítulos, incluindo o capítulo introdutório, tem-se:

- Capítulo 2: apresenta a revisão bibliográfica relacionada a caracterização mineral, seus principais métodos de caracterização e conceitos a respeito do sensoreamento hiperespectral, assim como abordagens e conceitos dos métodos de inteligência artificial e aprendizado de máquina.
- Capítulo 3: descreve os trabalhos relacionados ao tema e aplicações de imageamento hiperespectral na indústria.
- Capítulo 4: apresenta os materiais e métodos utilizados para a execução do projeto, subdividindo-se em métodos de inteligência artificial, estatísticos, desenvolvimento dos algoritmos e imageamento hiperespectral.
- Capítulo 5: exibe os resultados alcançados e discussões propostas.
- Capítulo 6: são dispostas a conclusão do projeto e considerações finais.

O texto é finalizado com a disposição das referências bibliográficas que auxiliaram o embasamento teórico ao decorrer da dissertação e disponibilizamos o código fonte dos algoritmos desenvolvidos.

2. Referencial Teórico e Fundamentação Científica

Neste capítulo é descrito o referencial teórico necessário para embasamento técnico na realização do projeto da dissertação. São expostos conceitos relacionados a caracterização mineralógica, imageamento remoto hiperespectral, inteligência artificial e métricas de avaliação.

2.1. Caracterização e conceitos sobre minérios e minerais

Com o intuito de otimizar o aproveitamento de recursos minerais, a caracterização de minérios se destaca como uma etapa essencial. Nesta etapa são fornecidos os elementos mineralógicos necessários para se dimensionar corretamente rotas de processos, assim como identificar ineficiências e perdas, otimizando, portanto, uma exploração em uma frente de lavra. A mineralogia e as propriedades inerentes ao minério, bem como os objetivos e a abrangência da caracterização, fazem com que existam variações no modo de se caracterizar uma amostra de minério. Processos tecnológicos adequados para um tipo de minério nem sempre são efetivos para um minério similar pois, de acordo com Anthony *et al.* (1990), os minérios apresentam características e peculiaridades próprias. Nos depósitos de minérios podem ocorrer as seguintes variações: (1) Composição mineralógica[1] devido à distribuição aleatória do mineral no depósito; (2) Granulometria[2] do mineral de interesse; (3) Relação dos minerais de ganga[3], entre outros.

A partir do resultado da interação de processos físico-químicos em ambientes geológicos, é formado um corpo natural sólido e cristalino denominado mineral (WILLS e FINCH, 2015). A classificação do mineral é definida por sua composição química e estrutura cristalina dos materiais que o compõem. Quando determinamos as proporções relativas dos diferentes elementos químicos de um mineral e a sua estrutura cristalina, obtemos sua composição. Em contrapartida, o minério se caracteriza como um agregado de minerais e ganga que é econômica e tecnologicamente viável para extração. Segundo Wills e Finch (2015), somente é possível classificar-se como minério quando este se concentra em quantidades elevadas e é possível de ser extraído da natureza. Os minérios de ferro são rochas a partir das quais pode ser obtido ferro metálico de modo economicamente viável. O ferro é encontrado sob a forma de óxidos, como a magnetita e a hematita.

[1]Devido à distribuição aleatória do mineral no depósito, a composição varia de acordo com cada ponto de. No nosso caso envolve Ferro, Sílica, Fosfato, Alumina, Manganês, Titânio, Magnésio e Carbonato.

[2]Granulometria é a determinação das dimensões das partículas do agregado e de suas respectivas porcentagens de ocorrência do mineral de interesse.

[3]Ganga são as impurezas sem valor encontradas juntos dos minérios de interesse econômico.

A Figura 2.1 apresenta exemplares de amostras de minérios de ferro da região de Itabira-MG e Mariana-MG. Ambos os minérios são Itabirito (hematita (Fe_2O_3) + quartzo (SiO_2)).

(a) Minério de ferro de Itabira-MG.

(b) Minério de ferro de Mariana-MG.

Figura 2.1: Amostras de minérios de ferro da região de Itabira-MG e Mariana-MG.

2.2. Caracterização mineralógica

No desenvolvimento de diferentes processos de beneficiamento que objetivam obter dados que serão úteis ao seu desenvolvimento e diminuir os erros operacionais, uma caracterização mineralógica de minério se faz essencial e não pode deixar de apresentar uma quantificação de seus constituintes representados pelo mineral de valor e pelos minerais pertencentes à ganga (HENLEY, 1983). Nesse tipo de análise encontramos a dificuldade de se manter a manutenção da representatividade, pois a quantidade da massa do material analisado é infinitamente menor em relação àquela encontrada em uma jazida. Para mantermos a representatividade em todas as etapas de avaliação de um minério, devemos manipular com cuidado a quantidade de minério analisada. Iniciando desde a amostragem na jazida até nos ensaios em escala de laboratório. Critérios de homogeneidade, conhecimento prévio do peso do mineral de interesse econômico e seu teor de ferro definem uma amostragem cuidadosa, levando em conta também uma possível indicação da granulometria de liberação do mineral valioso em relação ao elemento paragênese (BARBERY, 1991).

A identificação dos minerais baseia-se nas propriedades que o definem como um mineral, ou seja, propriedades físicas decorrentes, estrutura e composição química. Apesar das observações em escala mesoscópica (amostras de mão) serem bastante úteis, quando caracterizamos minerais é mais comum se trabalhar em escala microscópica pois assim é alcançada uma melhor identificação de boa parte dos minerais mais importantes. A Tabela 2.1 apresenta um comparativo com as principais vantagens, limitações e aplicações dos métodos de caracterização mineral que serão abordados no decorrer do texto.

Tabela 2.1: Comparativo dos métodos de caracterização mineral.

Método	Vantagens	Limitações	Aplicações
Microscopia Óptica	Precisos e flexíveis para análises qualitativas.	Ampliação e resolução limitadas.	Análise de amostras em grão.
Difratometria de Raios X	Alto poder de penetração no material, caráter não destrutivo.	Custo elevado.	Caracterização de materiais cristalinos.
Microscopia Eletrônica de varredura	Fácil preparação da amostra, ampla variedade de magnitude.	Dificuldade de examinar amostras isoladas e a impossibilidade em examinar amostras hidratadas.	Estudos de amostras mineralógicas ou petrográficas.
Espectroscopia no Infravermelho	Técnica rápida e de fácil execução. Método não destrutivo nem invasivo.	Alto investimento.	Medidas, controle de qualidade e análises dinâmicas.
Análises com Luz Ultravioleta	Obtenção de informações sobre crescimento de cristais ou inclusões que não se distinguem por outros métodos.	Uso restrito a minerais que apresentam fluorescência.	Teste da fluorescência de minerais.
Espectroscopia Mossbauer	Razão Fe^{2+} e Fe^{3+} Quantificação dos compostos de ferro, cristalinos e amorfos.	Tempo para aquisição de um espectro. Análise requer bom conhecimento da técnica.	Análise de campo. Substituição isomórfica (e.g. Fe^{2+} por Mg^{2+}, Ti^{2+}).
Espectroscopia Raman	Análise não destrutiva. Vasta gama de materiais podem ser analisados pelo método. Espectros são adquiridos em segundos.	Não pode ser usado para metais ou ligas. A detecção precisa de uma instrumentação sensível e altamente otimizada.	Identificação de rochas e minerais. Distribuição mineral e de fases nas seções rochosas.

2.2.1. Microscopia óptica

O trabalho em lupa ou microscópio estereoscópico permite analisar as amostras em grão. Segundo Gaspar *et al.* (1995), os minerais são identificados principalmente por cor, clivagem, traço, brilho e dureza. Para definirmos a cor de um mineral, este deve ser observado numa superfície à luz natural. Sua cor depende da absorção de certos comprimentos de onda do espectro solar incidente. A clivagem é uma propriedade física de um mineral que consiste na divisão segundo superfícies planas e brilhantes, de direções bem definidas e constantes. O traço é a cor apresentada por um mineral quando é reduzido a pó ou quando o utilizamos para riscar uma superfície. O brilho é o modo como o mineral reflete a luz natural. Os tipos mais comuns de brilho são metálico e não metálico. Finalmente, a dureza de um mineral é a resistência que ele oferece ao ser riscado.

De acordo com Klein e Dutrow (2009), os métodos de identificação de minerais mais comuns são as microscopias ópticas de luz transmitida para minerais transparentes, e de luz refletida para minerais opacos. Baseiam-se ambos na interação da luz (geralmente luz branca

do espectro visível) com os minerais e são precisos e flexíveis para análises qualitativas.

2.2.2. Difratometria de raios X

Segundo Melo *et al.* (2009), os raios X são radiações eletromagnéticas com comprimentos de onda que se estendem de 0,01 a 10 *nm*, podendo ser polarizados, refletidos e difratados. O feixe difratado sem mudança do comprimento de onda, resultante da dispersão dos raios X pelos elétrons dos átomos do cristal, é regido pela Equação 2.1, também conhecida por equação de Bragg.

$$n\lambda = 2d\text{sen}(\theta) \tag{2.1}$$

Na equação, o comprimento de onda dos raios no retículo refratado é representado por λ, a distância dos átomos é representada por d, n corresponde à ordem de difração e θ é o ângulo de Bragg que é o complemento do ângulo de incidência da óptica geométrica. A técnica de difração de raios X requer pequena quantidade de amostra ($< 1\ g$), além de ser um procedimento de baixo custo operacional. De acordo com Albers *et al.* (2002), os feixes de raios X são produzidos ao se bombardear o anodo por elétrons do catodo acelerados por alta voltagem. O feixe monocromático de raios X que incide sobre a amostra é difratado em cada plano cristalino que provoca uma interferência detectada pelo contador de radiação que é então armazenada em um registrador gráfico a partir do sinal eletrônico.

2.2.3. Microscópio eletrônico de varredura

Segundo Dedavid *et al.* (2007), existem dois princípios de microscopia eletrônica: a de transmissão e a de varredura. Na microscopia de transmissão, um feixe de elétrons atravessa a amostra e a imagem é projetada numa tela fluorescente, atingindo resolução de até 3 Å. Esta técnica permite a análise de defeitos e fases internas dos materiais. Na microscopia de varredura, o feixe de elétrons incide na amostra e os elétrons espalhados na superfície do material são captados, atingindo resoluções de 100 Å.

Ao utilizar o Microscópio Eletrônico de Varredura (MEV) é possível observar com detalhes as associações minerais, suas alterações, inclusões, zoneamentos e caracterizar os elementos químicos formadores do mineral (NEUMANN *et al.*, 2004). O MEV é o mais utilizado em tecnologia mineral, são utilizadas as imagens de elétrons onde o nível de cinza de cada pixel é proporcional ao peso atômico médio da fase naquele ponto.

A aplicabilidade do método depende de diversos fatores, inclusive do instrumental utilizado. As análises para se verificar proporções entre minerais principais, quando há bom contraste entre eles, são simples e rápidas. Quando o contraste é reduzido, já se torna necessário maior cuidado na calibração dos equipamentos, mais resolução nas imagens e melhores câmeras de vídeo e placas de interface (DUARTE *et al.*, 2003).

2.2.4. Espectroscopia no infravermelho

Segundo Stuart (2000), a espectroscopia no infravermelho é uma técnica analítica bastante útil na caracterização de substâncias químicas, pois fornece dados sobre a identidade e constituição estrutural de um composto ou sobre a composição qualitativa e quantitativa de misturas. Esta costuma ser uma técnica subestimada, pois além de fornecer informações complementares à difratometria de raios X, permite melhores identificações nos minerais de baixa cristalinidade, com altos índices de substituições no retículo ou materiais amorfos.

Para obtenção de espectros no infravermelho dos materiais sólidos, o método mais utilizado é o da pastilha com brometo de potássio prensada. Um espectro de infravermelho é composto de bandas de absorção intrinsecamente relacionadas aos movimentos moleculares, principalmente vibrações (DERRICK *et al.*, 2000).

2.2.5. Análises com luz ultravioleta

A fluorescência sob radiação ultravioleta é um método utilizado para identificar minerais. O uso da análise com luz ultravioleta é limitado pois apenas alguns minerais apresentam a propriedade de fluorescência, no entanto, é possível obter informações sobre crescimento de cristais que não seriam encontradas por outros métodos. A maior parte da fluorescência presente nos minerais é devida às impurezas.

Os elementos como manganês, urânio e terras raras induzem uma fluorescência nos minerais. Segundo Poole e Sims (2016), os minerais que sempre apresentam fluorescência são: scheelita, hidrozincita, willemita, autunita, malaquita, escapolita e fluorita. Entre os minerais que podem apresentar ou não fluorescência, dependendo dos ativadores, estariam incluídos: calcita, anglesita, wollastonita, nefelina, diamante e zirconita.

2.2.6. Espectroscopia Mossbauer

A espectroscopia Mossbauer, também conhecida por ressonância nuclear por emissão e absorção de raios gama sem recuo nuclear, permite a identificação de mais de 400 minerais que contem ferro (KLINGELHOEFER *et al.*, 2003). Quando se utiliza uma fonte radioativa de ^{57}Co em uma matriz, há uma emissão de raios gama causados por uma captura eletrônica. Com a captura eletrônica a fonte se transforma em ^{57}Fe e emite radiação gama. Essa radiação é absorvida por núcleos de ^{57}Fe na amostra analisada e é então reemitida. Essa absorção e remissão ocorrem sem perda energética (ressonância). Com a utilização do efeito Doppler (a fonte é posta para vibrar controladamente) (SCHRÖDER *et al.*, 2011) varre-se então um espectro de energias que permitem a distinção de pequenas variações de energia do núcleo (iterações hiperfinas) que resultam em espetros bem distintos de compostos ferrosos.

Os espectros são analisados por deconvolução e os parâmetros obtidos permitem a identificação dos diferentes compostos (DE SOUZA JR, 1999). A área de cada sub-espectro tem

relação com a quantidade de ferro no composto mineral em estudo. O tempo de aquisição de um espectro varia muito a depender da intensidade radioativa da fonte, da quantidade de ferro na amostra e da qualidade dos detectores utilizados.

2.2.7. Espectroscopia Raman

A espectroscopia Raman é definida como uma técnica de análise química não destrutiva capaz de fornecer informações detalhadas sobre a estrutura química, fase e polimorfia, cristalinidade e interações moleculares de uma amostra (FERRARO, 2003). Esta é uma técnica dentro da espectroscopia vibracional, baseada na dispersão inelástica da luz. Segundo, Mitsutake *et al.* (2019), desde o desenvolvimento do primeiro espectrômetro Raman comercial em 1953, os avanços nos lasers e detectores e a descoberta de novos fenômenos expandiram o uso dessa técnica em vários campos de pesquisa.

Segundo Karr (2013), a quantificação de proporções minerais em rochas e solos por espectroscopia Raman em uma superfície é melhor feita com muitos espectros de feixe estreito de diferentes locais da rocha ou do solo (representatividade mineral). A proporção de cada mineral na rocha ou no solo pode ser determinada a partir da fração dos espectros que contêm seus picos.

De acordo com Colthup (2012), um espectro Raman apresenta vários picos, mostrando a intensidade e a posição do comprimento de onda da luz dispersa. Geralmente, o espectro é uma impressão digital química distinta para uma molécula ou material específico e pode ser usado para identificar rapidamente o material ou distingui-lo dos outros. As bibliotecas espectrais Raman são frequentemente usadas para identificação de um material com base em seu espectro Raman, bibliotecas contendo milhares de espectros são rapidamente pesquisadas para encontrar uma correspondência com o espectro analisado (MCCREERY, 2005).

2.3. Sensoriamento remoto

Segundo Gupta (2017), sensoriamento remoto é definido como o conjunto de técnicas que possibilitam obter informações relevantes sobre alvos na superfície terrestre como objetos, áreas e fenômenos. A obtenção desses dados é possível devido à interação da radiação eletromagnética com a superfície analisada. É comum encontrarmos exemplos desses sensores em satélites, aviões, drones e a nível de campo.

A intensidade do espectro eletromagnético é medida por sensores remotos, a partir desta podemos obter imagens nas regiões visíveis a olho nu até a região do infravermelho, medindo a intensidade da radiação eletromagnética refletida ou emitida, a depender da região do espectro eletromagnético analisada (TOTH e JÓŹKÓW, 2016). Um fator preponderante nessa ciência é a resolução dos sensores. O conceito de resolução se divide em 4 classes descritas a seguir: espacial, espectral, radiométrica e temporal.

- **Resolução Espacial:** está relacionada com a capacidade do sensor de dividir os elementos na superfície observada. Quanto maior a resolução espacial, maior o nível de detalhe observado.

- **Resolução Espectral:** consiste na propriedade de um sensor operar em várias bandas espectrais. Os sensores que operam em centenas de bandas são conhecidos como hiperespectrais.

- **Resolução Radiométrica:** possui relação com o nível de quantização ou sensibilidade de um sensor em detectar pequenas diferenças na energia eletromagnética, expressa pelo número de tons de cinza.

- **Resolução Temporal:** sua definição é dada em função do tempo de revisita de um sensor para um mesmo ponto de uma superfície analisada.

2.3.1. Resolução espectral

Sistemas de sensoriamento remoto registram energia em várias faixas de comprimento de onda separadas em diversas resoluções espectrais. Sistemas multiespectrais fazem referência a sensores com poucas bandas, por exemplo 7, 16 bandas. Os sensores hiperespectrais detectam centenas de bandas espectrais muito estreitas em todas as porções do visível ao olho humano, do infravermelho próximo e infravermelho médio do espectro eletromagnético. Sua alta resolução espectral facilita a discriminação fina entre diferentes alvos com base em sua resposta espectral em cada uma das bandas (LANDGREBE, 2005).

Segundo Rees e Pellika (2010), a resolução espectral descreve a capacidade de um sensor para definir intervalos de comprimento de onda. Quanto melhor for a resolução espectral, mais estreita será a faixa de comprimento de onda para uma determinada banda.

2.3.2. Espectroscopia de imageamento

A imagem hiperespectral, por meio de imagens feitas a partir de informações espectrais coletadas por um espectrômetro pode ilustrar a composição química dependendo da região a ser analisada, podendo ser inferida alguma informação relacionada com a composição química e/ou cristalinidade do material. Porém, sempre são necessárias associações com análises químicas do material. A imagem hiperespectral pode ser feita por meio das espectroscopias: Visible and Near Infrared (VNIR) entre 400 - 1.000 *nm*, Near Infrared (NIR) entre 900 - 1.700 *nm*, Short-wavelength Infrared (SWIR) entre 1.000 - 2.500 *nm*, Mid-wavelength Infrared (MWIR) entre 2,7 – 5,3 μm, Long-wavelength Infrared (LWIR) entre 8 - 12,4 μm, dentre outras. É uma ferramenta de alto potencial para inúmeras aplicações em indústrias, pesquisas e agricultura, por meio de análises não invasivas. Exemplos dessas aplicações podem ser encontrados em Dong *et al.* (2019), Gowen *et al.* (2007) e Haboudane *et al.* (2004).

2.4. Inteligência Artificial

Segundo Dunjko e Briegel (2018), a Inteligência Artificial (IA) é um campo da computação que estuda a construção de entidades inteligentes, ou seja, máquinas que parecem ter inteligência humana. Ainda, de acordo com Nilsson (2014), a Inteligência Artificial visa o entendimento da inteligência humana para realizar sua reprodução na computação. IA é um ramo da Ciência da Computação que se iniciou com a intenção de reproduzir modelos cognitivos em computador.

O aprendizado de máquina é uma forma de IA onde um algoritmo computacional constrói, a partir de dados, modelos de aprendizado para a resolução de problemas (KUMAR *et al.*, 2017). Seu objetivo é criar modelos de software que são treinados com grandes volumes de dados e usados para prever resultados, tendências e padrões (WATT *et al.*, 2020). Alguns exemplos de tarefas resolvidas com maior eficiência pelo aprendizado de máquina incluem:

- **Reconhecimento de padrões:** objetos em cenas reais, identidades ou expressões faciais, palavras escritas ou faladas;
- **Detecção de anomalias:** sequências incomuns de transações de cartão de crédito e padrões incomuns de leituras de sensores em máquinas de uma indústria têxtil;
- **Previsão:** preços de ações ou taxas de câmbio, quais filmes uma pessoa gostaria de assistir, previsão de vendas, etc.

Neste estudo, compara-se o desempenho de dois Algoritmos de Aprendizado de Máquina (AAP): Redes Neurais Artificiais (RNA) e Random Forest (RF). Esses AAPs foram escolhidos devido à sua grande presença na literatura de problemas de classificação e regressão (DOMINGOS, 2015), e ao fato de representarem diferentes abordagens de aprendizado de máquina, o que garante uma diversidade algorítmica de aprendizado de máquina. Nesta seção, os conceitos básicos dos AAPs mencionados são explicados, assim como são expostos os conceitos de *overfitting* (sobreajuste ou superajuste) e *underfitting* (sub-ajuste) que em aprendizado de máquina são classificações ou conceitos a respeito do ajuste de dados dos modelos.

2.4.1. Ajustes de modelos

Segundo Brownlee (2016), os algoritmos de aprendizado de máquina realizam o ajuste do modelo (do inglês, *model fit*), que ocorre enquanto ele está sendo treinado com base nos dados para que se torne possível realizar previsões (do inglês, *model predict*). Essa compreensão orientará a tomar medidas corretivas. Pode-se determinar se um modelo preditivo está fazendo o sub-ajuste ou o sobreajuste dos dados de treinamento consultando o erro de previsão nos dados de treinamento e nos dados de avaliação. Compreender o ajuste de modelo é importante para entender a causa raiz da precisão insatisfatória do modelo (JABBAR e KHAN, 2015).

O *overfitting* acontece quando o modelo aprende os detalhes e o ruído dos dados de treinamento na medida em que afeta negativamente o desempenho do modelo em novos dados. Isso significa que o ruído ou flutuações aleatórias nos dados de treinamento são captados e aprendidos como conceitos pelo modelo. Os conceitos aprendidos não se aplicam a novos dados de entrada e portanto afetam negativamente a capacidade de generalização dos modelos. (BROWNLEE, 2016).

O *underfitting*, em contrapartida ao *overfitting*, acontece quando um modelo de aprendizado de máquina não é complexo o suficiente para capturar com precisão as relações entre os atributos de um conjunto de dados e uma variável de destino. Um modelo mal equipado resulta em resultados problemáticos ou errôneos em novos dados, ou dados nos quais não foi treinado, e geralmente apresenta desempenho ruim, mesmo em dados de treinamento (BROWNLEE, 2016).

Na Figura 2.2 são dispostas representações gráficas do modelo *underfitting* e *overfitting* de acordo com o que foi explicado anteriormente e um exemplo genérico de um caso de ajuste ideal do modelo de dados.

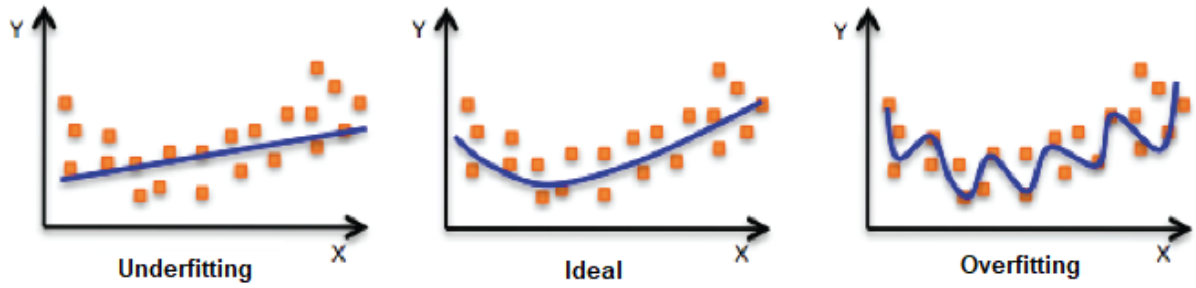


Figura 2.2: Representação dos modelos de ajuste de dados de modelos (BROWNLEE, 2016).

2.4.2. Redes neurais artificiais

Redes Neurais Artificiais (RNA) são definidas como um modelo matemático para a simulação de uma rede de neurônios biológicos, ou seja, são as partes de um sistema de computação projetado para simular a maneira como o cérebro humano analisa e processa as informações aprendidas. Em termos simples, é uma ferramenta de computação que pode aprender padrões e prever resultados (ZHAO *et al.*, 2012).

Segundo Shanmuganathan (2016), a forma mais simples de uma RNA é o Multilayer Perceptron (MLP), que consiste em três camadas: a camada de entrada, a camada oculta e a camada de saída. Um Perceptron é um classificador linear, ou seja, é um algoritmo que classifica a entrada separando duas categorias com uma linha reta. Um Perceptron produz uma única saída com base em várias entradas de valor real, formando uma combinação linear usando os pesos (e às vezes passando a saída através de uma função de ativação não linear).

Um Multilayer Perceptron (MLP) é uma rede neural artificial composta por mais de um Perceptron. Sua composição consiste em uma camada de entrada que recebe os dados, uma

camada de saída que toma decisões sobre a entrada, e entre essas camadas, um número de camadas ocultas. MLPs com uma camada oculta são capazes de aproximar qualquer função contínua (SHANMUGANATHAN, 2016).

Haykin *et al.* (2009) afirmam que os processos de aprendizagem de uma rede neural artificial são determinados por como as mudanças de parâmetro ocorrem. Assim, o processo de aprendizagem de uma RNA é dividido em três partes:

1. A estimulação por extração de exemplos de um ambiente;
2. A modificação de seus pesos através de processos iterativos a fim de minimizar o erro de saída da RNA;
3. A rede responde de uma nova maneira como resultado das mudanças ocorridas.

Denomina-se algoritmo de aprendizagem um conjunto de regras bem definidas para a solução de um problema de aprendizagem (GOMES, 2011). Um importante fator é a maneira pela qual uma rede neural se relaciona com o ambiente. Nesse contexto, segundo Alpaydin (2016), existem os seguintes paradigmas de aprendizagem:

- **Aprendizagem Supervisionada:** é a tarefa de encontrar uma função a partir de dados de treinamento rotulados. O objetivo é encontrar os parâmetros ótimos que ajustem um modelo que possa prever rótulos desconhecidos em outros objetos (o conjunto de teste);
- **Aprendizagem Não Supervisionada:** é aquele que para fazer modificações nos valores das conexões sinápticas não se usam as informações sobre a resposta da rede, isto é, se a resposta está correta ou não. Neste caso, usa-se outro esquema, tal que, para exemplos semelhantes, a rede responda de modo semelhante;
- **Aprendizagem por Reforço:** é o treinamento de modelos de aprendizado de máquina para tomar uma sequência de decisões. O agente aprende a atingir uma meta em um ambiente incerto e potencialmente complexo. Na aprendizado por reforço, o sistema de inteligência artificial enfrenta uma situação, o computador utiliza tentativa e erro para encontrar uma solução para o problema. Para que a máquina faça o que o programador deseja, a inteligência artificial recebe recompensas ou penalidades pelas ações que executa. Seu objetivo é maximizar a recompensa total.

A configuração de parâmetros impacta diretamente no processo de aprendizagem de uma RNA. Alguns exemplos de parâmetros são: taxa de aprendizado, taxa de *momentum*, critérios de parada e forma de treinamento em rede. A Figura 2.3 mostra um exemplo de uma topologia MLP totalmente conectada, com camadas de entrada, ocultas e de saída.

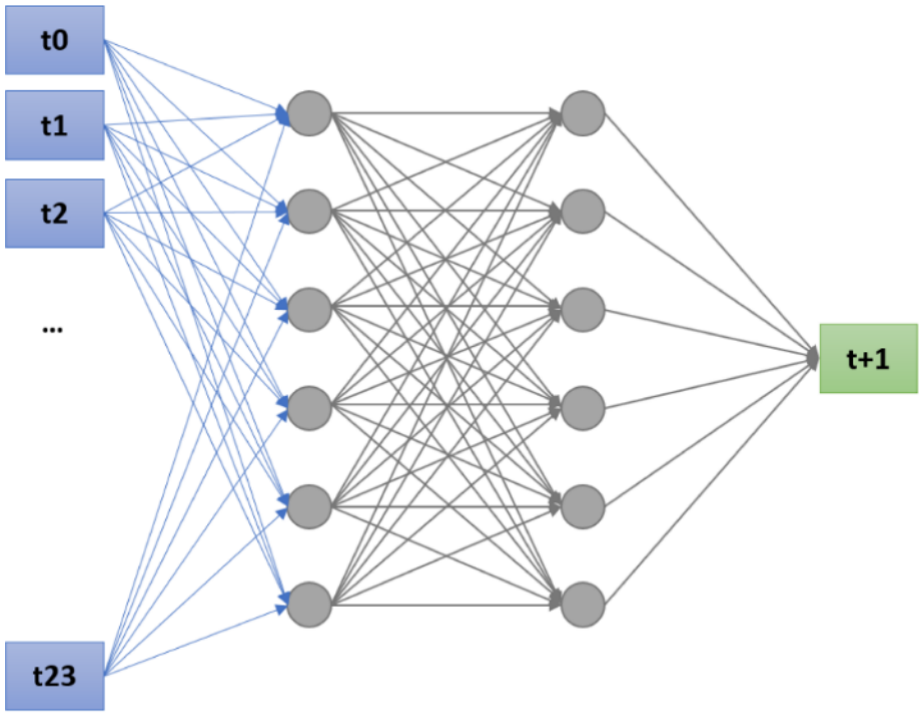


Figura 2.3: Exemplo de esquema de uma RNA Multilayer Perceptron com duas camadas ocultas com 6 neurônios cada (SANTOS *et al.*, 2019).

2.4.3. Random forest

Segundo Cutler *et al.* (2012), Random Forest (RF) são conjuntos de árvores de decisão que votam juntos em uma classificação. Cada árvore é construída por acaso e seleciona um subconjunto de recursos aleatoriamente de um subconjunto de pontos de dados. A árvore é então treinada nesses pontos de dados (somente nas características selecionadas), e o restante *fora do cesto* é usado para avaliar a árvore. As árvores de decisão estabelecem regras para tomada de decisão. É criada uma estrutura similar a um fluxograma, com nós onde condições são verificadas, e se atendidas o fluxo segue por um ramo, caso contrário, pelo outro, sempre levando ao próximo nó, até a finalização da árvore (BIAU e SCORNET, 2016). Com os dados de treino, o algoritmo busca as melhores condições e onde inserir cada uma dentro do fluxo. Conforme proposto por Breiman (2001), as características da árvore de decisão são: (i) facilidade de implementação; (ii) boas propriedades de generalização; (iii) o algoritmo gera mais informações do que apenas rótulo de classe; (iv) funciona eficientemente em grandes bases de dados; (v) pode lidar com milhares de variáveis de entrada sem eliminação variável; e (vi) fornece estimativas de quais variáveis são importantes na estimação.

Uma das vantagens do Random Forest é que pode ser utilizado tanto para regressão quanto para classificação e é fácil visualizar a importância relativa que ele atribui para cada característica na suas entradas (AO *et al.*, 2019). Porém, um dos grande problemas em aprendi-

zagem de máquina é o sobreajuste (overfitting), mas a maior parte do tempo isto não ocorrerá tão facilmente com um Random Forest qualquer. Isto porque, se há árvores suficientes, o Random Forest não irá sobreajustar o modelo (HORNING *et al.*, 2010).

Segundo Rodriguez-Galiano *et al.* (2012), a maior limitação do Random Forest é que uma quantidade grande de árvores pode tornar o algoritmo lento e ineficiente para predições em tempo real. Em geral, estes algoritmos são rápidos para treinar, mas muito lentos para fazer predições depois de treinados. Uma predição com mais acurácia requer mais árvores, o que faz o modelo ficar mais lento. Em muitas aplicações do mundo real o Random Forest é rápido o suficiente, mas pode certamente haver situações onde a performance em tempo de execução é importante e outras abordagens são mais apropriadas. Uma visão esquemática de um Random Forest, que é uma série de árvores de decisão com um mecanismo conjunto - geralmente voto de maioria para classificação ou média de saídas para regressão, pode ser vista na Figura 2.4.

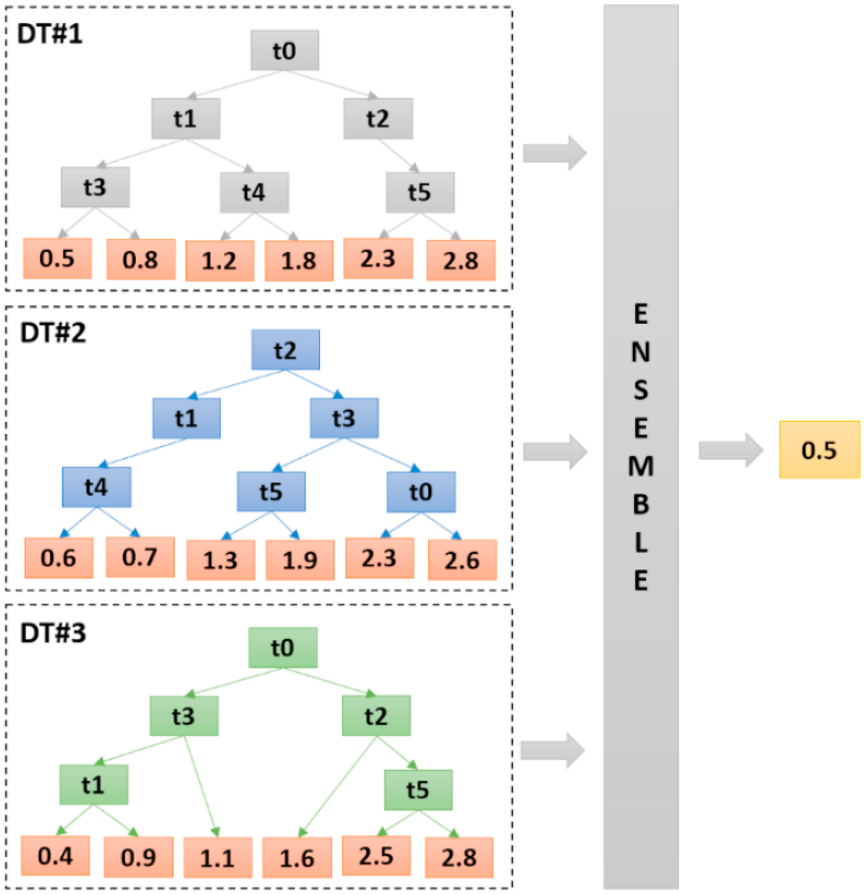


Figura 2.4: Exemplo de esquema de uma Random Forest (SANTOS *et al.*, 2019).

2.4.4. Seleção de variáveis

Geralmente, na análise de dados, temos centenas ou até milhões de variáveis e queremos uma maneira de criar um modelo que inclua apenas os recursos mais importantes. Isso tem três benefícios. Primeiro, tornamos nosso modelo mais simples de interpretar. Segundo, podemos reduzir a variação do modelo. Finalmente, podemos reduzir o custo (e o tempo) computacional do treinamento de um modelo. O processo de identificação apenas dos recursos mais relevantes é chamado de *seleção de variáveis*.

O conhecimento da importância das variáveis indicadas pelos modelos de aprendizado de máquina podem beneficiar de várias maneiras, por exemplo:

- Melhor entendimento da lógica do modelo e também trabalhar em sua melhor modelagem, concentrando-se apenas nas variáveis importantes;
- Remoção de *n* variáveis que não são tão significativas e têm desempenho semelhante ou melhor em um tempo de treinamento muito menor.

Segundo Zhou *et al.* (2016), um dos principais tópicos no desenvolvimento de modelos preditivos é a identificação de variáveis que são preditores de um determinado resultado. Métodos automatizados de seleção de modelo, como regressão linear, são soluções clássicas para esse problema, mas geralmente são baseados em fortes suposições sobre a forma funcional do modelo ou a distribuição de resíduos. Neste contexto, é proposto um método de seleção alternativo baseado na técnica de Random Forest.

A principal limitação do uso de um conjunto de Random Forest reside em seu poder explicativo: as previsões são o resultado de uma caixa preta na qual é impossível distinguir a contribuição dos preditores únicos. Com a RF, esse problema é ainda mais crucial, porque o método funciona muito bem, especialmente na presença de um pequeno número de preditores informativos ocultos entre um grande número de variáveis de ruído (MENZE *et al.*, 2009).

Para superar essa limitação, são dispostos dois métodos simples para a seleção de variáveis:

- **Diminuição da impureza média:** cada nó nas RFs é uma condição em um único recurso, projetado para dividir o conjunto de dados em dois, para que valores de resposta semelhantes terminem no mesmo conjunto. A medida com base na qual a condição ótima é escolhida é chamada de impureza. Para classificação, normalmente é a impureza de Gini ou ganho/entropia de informações e, para as árvores de regressão, é a variação. Assim, ao treinar uma árvore, pode-se calcular quanto cada recurso diminui a impureza ponderada em uma árvore. Para uma RF, a diminuição da impureza de cada recurso pode ser calculada como média e os recursos são classificados de acordo com esta medida (LI *et al.*, 2017).

- **Diminuição da precisão média:** Em primeiro lugar, a seleção de recursos com base na redução de impurezas é influenciada pela preferência por variáveis com mais categorias. Em segundo lugar, quando o conjunto de dados possui dois (ou mais) recursos correlatos, do ponto de vista do modelo, qualquer um desses recursos correlatos pode ser usado como preditor, sem preferência concreta de um sobre os outros (CAI *et al.*, 2018). Mas uma vez que um deles é usado, a importância de outros é significativamente reduzida, pois efetivamente a impureza que eles podem remover já é retirada pelo primeiro recurso. Como consequência, eles terão uma menor importância relatada. Esse não é um problema quando queremos usar a seleção de recursos para reduzir o excesso de ajustes, pois faz sentido remover recursos que são na maioria das vezes duplicados por outros recursos. Porém, ao interpretar os dados, isso pode levar à conclusão incorreta de que uma das variáveis é um forte preditor enquanto as outras do mesmo grupo não são importantes, enquanto na verdade elas são muito próximas em termos de relacionamento com a variável de resposta (ALELYANI *et al.*, 2018).

2.5. Métricas de avaliação

A estatística descritiva é uma forma de análise matemática que utiliza modelos quantificados, representações e sinopses para um determinado conjunto de dados experimentais. Esta é uma parte da matemática aplicada que fornece dados para a coleta, organização, descrição, análise e a interpretação de dados para a utilização dos mesmos na tomada de decisão (MARÔCO, 2018).

Neste trabalho faremos uso das seguintes métricas para obtermos relações entre as variáveis e observamos os erros das estimações dos teores de ferro: o Coeficiente de Determinção (R^2), o Erro Médio Absoluto (MAE) e o Erro Médio Quadrático (MSE).

O **Coeficiente de Determinção**, também chamado de R^2, é uma medida de ajustamento de um modelo estatístico linear generalizado, como a regressão linear, em relação aos valores observados (FIGUEIREDO FILHO e SILVA JÚNIOR, 2009). O R^2 varia entre 0 e 1, indicando, em porcentagem, o quanto o modelo consegue explicar os valores observados. Quanto maior o R^2, mais explicativo é o modelo, melhor ele se ajusta à amostra.

O **Erro Médio Absoluto** é uma medida da diferença entre duas variáveis contínuas. O MAE mede a magnitude média dos erros em um conjunto de previsões. É a média da amostra de teste das diferenças absolutas entre a previsão e a observação real, em que todas as diferenças individuais têm peso igual (WILLMOTT e MATSUURA, 2005).

Segundo Gupta *et al.* (2009), o **Erro Médio Quadrático** é definido como sendo a média da diferença entre o valor do estimador e do parâmetro ao quadrado. O MSE é um dos dois principais indicadores de desempenho para um modelo de previsão de regressão. Ele mede a precisão do modelo de previsão calculando a média dos quadrados das diferenças entre os valores previstos e reais.

A Tabela 2.2 apresen as vantagens e limitações dos métodos estatísticos considerados para o escopo desta dissertação.

Tabela 2.2: Comparativo das métricas de avaliação.

Método	Vantagens	Limitações
Coeficiente de Determinação (R^2)	O valor é independente de qualquer unidade usada para medir as variáveis.	É necessário que as duas variáveis sejam medidas em um nível quantitativo contínuo.
Erro Médio Absoluto (MAE)	Garante o uso do valor absoluto nos cálculos matemáticos e implicações mais fáceis de entender.	Não trata erros isolados acima do valor real de modo diferente dos que estão abaixo dele.
Erro Médio Quadrático (MSE)	Benefício de penalizar grandes erros, por isso pode ser mais apropriado quando erros grandes são indesejáveis.	Não é expresso na mesma unidade dos dados originais.

2.6. Considerações

Neste capítulo, foi apresentada uma breve revisão dos conceitos de mineralogia e caracterização mineral, as seguintes técnicas de aprendizado de máquina investigadas neste trabalho: MLP e RF; assim como um resumo sobre as métricas estatísticas relevantes para o projeto. Selecionamos essas técnicas por seus excelentes desempenhos e popularidade. A seguir, colocaremos em prática cada uma dessas técnicas a fim de comparar seus desempenhos na tarefa de estimar o teor de ferro em amostras de minério de ferro.

3. Trabalhos Relacionados

Este capítulo descreve resumidamente alguns trabalhos recentes com o tema imageamento hiperespectral em conjunto de técnicas de aprendizado de máquina encontrados na literatura, assim como analisa algumas patentes depositadas referentes as tecnologias citadas. Entretanto, a análise técnica destas soluções não é viável pois seus métodos de coleta e modelos de análise de dados não estão publicamente disponíveis.

3.1. Aplicações de imageamento hiperespectral na indústria

O trabalho de Zhu *et al.* (2017) investiga a viabilidade e potencialidade da detecção pré-sintomática da doença do tabaco usando imagens hiperespectrais, combinadas com o método de seleção de variáveis e os classificadores de aprendizado de máquina. Imagens de folhas saudáveis e infectadas foram adquiridas por um sistema de imagem de refletância hiperespectral, cobrindo a faixa espectral de 380 - 1.023 nm. Além disso, diferentes algoritmos de aprendizado de máquina foram desenvolvidos e comparados para detectar e classificar estágios da doença, recursos de textura e fusão de dados, respectivamente. O desempenho de modelos manifestou resultados com precisão de classificação de calibração e previsão acima de 80%; as precisões eram de até 95%, empregando Modelos de Rede Neural de Retropropagação (BPNN), Máquina de Aprendizado Extremo (ELM) e Modelos de Máquina de Vetores de Suporte de Mínimos Quadrados (LS-SVM).

Gewali *et al.* (2018) analisam e comparam métodos recentes de análise de imagens hiperespectrais baseados em aprendizado de máquina publicados na literatura. Organizaram os métodos pela tarefa de análise de imagem e pelo tipo de algoritmo de aprendizado de máquina, e apresentaram um mapeamento bidirecional entre as tarefas de análise de imagem e os tipos de algoritmos de aprendizado de máquina que podem ser aplicados a eles. O artigo é abrangente na cobertura de tarefas de análise de imagens hiperespectrais e algoritmos de aprendizado de máquina. As tarefas de análise de imagem consideradas são classificação da cobertura do solo, detecção de alvos, desmistificação e estimativa de parâmetros físicos. Os algoritmos de aprendizado de máquina abordados são modelos gaussianos, regressão linear, regressão logística, máquinas de vetores de suporte, modelo de mistura gaussiana, modelos lineares latentes, modelos lineares esparsos, modelos de mistura gaussiana, aprendizado de conjuntos, modelos gráficos direcionados, modelos gráficos não direcionados, agrupamento, processos gaussianos, Processos de Dirichlet e aprendizado profundo.

As observações espectrais ao longo do espectro em muitas bandas espectrais estreitas através de imagens hiperespectrais fornecem informações valiosas para o reconhecimento de materiais e objetos, que podem ser consideradas como uma tarefa de classificação. A maioria dos estudos e esforços de pesquisa existentes segue o paradigma convencional de reconhecimento de padrões, que se baseia na construção de recursos artesanais complexos. No entanto,

raramente se sabe quais recursos são importantes para o problema em questão. Em contraste com essas abordagens, (MAKANTASIS *et al.*, 2015) propuseram um método de classificação baseado em aprendizado profundo que constrói hierarquicamente recursos de alto nível de maneira automatizada. O método explora uma Rede Neural Convolucional para codificar informações espectrais e espaciais dos pixeis e um Perceptron de várias camadas para conduzir a tarefa de classificação. Foram apresentados resultados experimentais e validação quantitativa em conjuntos de dados amplamente utilizados, mostrando o potencial da abordagem desenvolvida para a classificação precisa de dados hiperespectrais.

A pesquisa de Sandino *et al.* (2018) propõe uma estrutura que consolida informações baseadas em locais e recursos de sensoriamento remoto para detectar e segmentar deteriorações por patógenos fúngicos em florestas naturais e de plantações. Esta abordagem é ilustrada com um caso de experimentação de ferrugem em árvores de chá de papel em New South Wales, Austrália. O método integra veículos aéreos não tripulados (UAVs), sensores de imagem hiperespectrais e algoritmos de processamento de dados usando o aprendizado de máquina. As imagens são adquiridas usando uma câmera Headwall Nano-Hyperspec®, e processadas na linguagem de programação Python usando o eXtreme Gradient Boosting (XGBoost), a Biblioteca de Abstração de Dados Geoespaciais (GDAL) e as bibliotecas Scikit-learn. No total, 11.385 amostras foram extraídas e rotuladas em cinco classes: duas classes para status de deterioração e três classes para objetos de segundo plano. As informações revelam taxas de detecção individual de 95% para árvores saudáveis, 97% para árvores deterioradas e uma taxa global de detecção de várias classes de 97%. A metodologia é versátil para ser aplicada a conjuntos de dados adicionais obtidos com diferentes sensores de imagem e ao processamento de grandes conjuntos de dados com ferramentas de freeware.

O crescimento econômico do país depende principalmente das fontes minerais e de energia. Nos últimos anos, há uma crescente pressão para reduzir o impacto ambiental e social por meio da exploração mineral. Os dados do satélite de sensoriamento remoto desempenham um papel vital e são capazes de detectar recursos minerais. O sensoriamento remoto hiperespectral é uma ferramenta eficaz para aplicações de exploração mineral, pois fornece desenvolvimento significativo nas últimas três décadas. No artigo de Sudharsan *et al.* (2019) é fornecida uma revisão atualizada e focada das técnicas de mapeamento mineral usando a imagem hiperespectral. Um fator chave para o sucesso da imagem hiperespectral é a resolução espacial, espectral e temporal e a capacidade de cobrir grandes superfícies da terra usando as modernas tecnologias de sensores. Esta revisão se concentra principalmente nos fundamentos das técnicas de imagem hiperespectral e técnicas de Imagem Analítica e Geofísica (AIG) para mapeamento mineral. Esta revisão pode ser uma linha de base útil para pesquisas futuras em exploração mineral usando análise de imagem hiperespectral.

3.2. Análise de patentes

Esta análise foi realizada no intuito obter informações a respeito do perfil de desenvolvimento tecnológico do segmento tendo o auxílio de informações como as mais recentes invenções e o tipo de tecnologia e metodologia utilizada para alcançar resultados semelhantes ao proposto pelo presente trabalho.

Começamos analisando a patente: CN110031414A - Multilayer perceptron hyperspectral mineral classification method based on spectral absorption index publicada em Junho de 2019 e depositada na China. O aplicante é UNIV BEIHANG e os inventores são: DENG KEWANG, LI NA e ZHAO HUIJIE. A invenção divulga um método de classificação de minerais utilizando imagens hiperespectrais e um modelo perceptron multicamadas com base em um índice de absorção espectral. O método de classificação compreende as seguintes etapas: (1) os dados hiperespectrais são lidos; (2) as espécies e o número de minerais são determinados e as amostras de treinamento e de teste são selecionadas; (3) um vetor de índice de absorção espectral é determinado de acordo com uma banda de onda característica mineral; (4) um modelo de rede de profundidade baseado no perceptron multicamadas é estabelecido; (5) os parâmetros do modelo são treinados e uma estratégia anti-sobreajuste é construída; e (6) as imagens hiperespectrais são classificadas e um mapa de classificação mineral é obtido. De acordo com o método de classificação de minerais hiperespectrais perceptron multicamadas, um modelo perceptron multicamadas é usado como base do modelo, o vetor de índice de absorção espectral mineral é usado como entrada e a classificação mineral com base nos dados hiperespectrais é alcançada (DENG KEWANG, CN Patent 110031414A, Jun. 2019).

Em seguida fazemos a análise da patente: CN107145830A - Hyperspectral image classification method based on spatial information enhancement and deep belief network publicada em Setembro de 2017 e depositada na China. O aplicante é UNIV XIDIAN e os inventores são LI JIAOJIAO, LI YUNSONG e SUN LIPING. A invenção divulga um método de classificação de imagem hiperespectral com base no aprimoramento de informação espacial e uma rede de aprendizado profundo. O método compreende as seguintes etapas: 1) realização de normalização e seleção de banda em uma imagem hiperespectral para obter uma imagem com um valor de refletância de 0 a 1; 2) realizar o aumento da informação espacial na imagem hiperespectral por meio de agrupamento de bandas e filtragem de solo e guiada; 3) construir um modelo de rede de aprendizado profundo de acordo com as características da imagem hiperespectral após o aumento da informação espacial; e 4) realizar o treinamento do modelo na imagem hiperespectral após o aumento da informação espacial e utilizar o modelo obtido para realizar a predição da categoria a fim de obter um resultado de classificação. De acordo com o método, a informação espacial da imagem hiperespectral é efetivamente aprimorada sem perder a informação espectral, a precisão da classificação é melhorada e o método pode ser usado para meio ambiente, clima, agricultura e exploração mineral (LI JIAOJIAO, CN Patent 107145830A, Set. 2017).

Por fim, é apresentada a patente CN109283148A - Method for automatically identifying rock minerals based on spectral information publicada em Janeiro de 2019 e depositada na china. O aplicante é BEIJING RES INSTITUTE OF URANIUM GEOLOGY e os inventores são LIU HONGCHENG, MENG SHU, QIU JUNTING, WANG JIANGANG, YE FAWANG e ZHANG CHUAN. A invenção pertence ao campo da aplicação de sensoriamento remoto hiperespectral e divulga um método para a identificação de minerais de rocha com base em informações espectrais. O método compreende as seguintes etapas: 1) varredura de minerais de rocha a serem classificados usando um sensor hiperespectral para obter dados espectrais minerais; 2) armazenar os dados obtidos em um banco de dados; 3) extrair uma parte dos dados do banco de dados; 4) identificar informações de espécies minerais correspondentes aos dados extraídos de acordo com as características usando interpretação manual e armazenar os dados espectrais extraídos; 5) estabelecer um sistema de aprendizagem de inteligência artificial e realizar aprendizagem e treinamento usando a biblioteca de amostra de aprendizagem; 6) detectar os dados espectrais no banco de dados a serem detectados usando o sistema de aprendizagem de inteligência artificial após o treinamento e otimização de aprendizagem para identificar as informações de espécies minerais; 7) armazenar um resultado de identificação em um banco de dados de resultados de identificação. Com a adoção do método, o consumo de recursos humanos pode ser reduzido, o grau de automação de trabalho é melhorado, a eficiência de processamento de dados de varredura espectral é melhorada e a eficiência econômica é melhorada (LIU HONGCHENG, CN Patent 109283148A, Jan. 2019).

Estas patentes foram selecionadas pois são as que mais se assemelham às metodologias e técnicas do presente trabalho. Conseguindo assim situar o nível tecnológico e de desenvolvimento de outros projetos relacionados ao tema. Nesta busca não foram encontradas patentes relevantes ao tema depositadas no Brasil.

4. Materiais e Métodos

Neste capítulo estão descritos os materiais e métodos que foram utilizados no decorrer da pesquisa para que fossem alcançados os resultados a serem expostos no capítulo seguinte. Aqui são apresentadas as etapas de programação e desenvolvimento dos algoritmos, o imageamento hiperespectral das amostras e o tratamento dos dados obtidos.

4.1. Formação da base de dados

4.1.1. Sensor hiperespectral

O conjunto de dados da espectroscopia de imagem foi adquirido pelo sensor hiperespectral Specim FX10 acoplado em uma plataforma de laboratório na altura de 30 cm. O sensor possui 224 bandas espectrais que cobrem comprimentos de onda na região do VNIR entre 400 - 1.000 *nm*, com 5,5 *nm* de amostragem espectral, 1.024 pixeis espaciais e um campo de visão de 38°. O Spectralon® (Labsphere, Inc) foi utilizado como material de referência, considerando 100% de refletância na faixa espectral de interesse (400 - 1.000 *nm*).

A análise espectral do minério de ferro foi realizada para identificar as características de absorção de cada litologia, especialmente aquelas associadas ao ferro férrico e ferroso. Os espectros removidos a vácuo calculados para a faixa de 400 - 1.000 *nm* foram usados para destacar as bandas de absorção. As características de absorção de ferro estão relacionadas principalmente à hematita e goethita.

Para posterior realização de análises através de técnicas de aprendizado de máquina, foram selecionados 24.000 pixeis de cada amostra. 95% dos pixeis das amostras foram coletados e desconsiderados os 5% dos pixeis que ficam mais na borda da placa petri. Após a captura das imagens é necessário passar por um pré-processamento para obter a imagem em refletância, realizada por software computacional. Em seguida são gerados dados espectrais que possibilitam montar uma biblioteca espectral para caracterização dos minérios a partir de características em suas curvas.

4.1.2. Amostras

As amostras investigadas são exibidas na Figura 4.1. Elas foram coletadas considerando várias litologias de uma frente de minério na Mina de Brucutu (São Gonçalo do Rio Abaixo, Minas Gerais - Brasil). As amostras apresentam uma ampla distribuição do conteúdo de Fe que variam de 37% a 62%. Informações relacionadas ao conteúdo global de ferro das amostras de minério de ferro coletadas estão disponíveis na Tabela 4.1.

Figura 4.1: Amostras de minério de ferro.

Tabela 4.1: Conteúdo global de ferro das amostras de minério de ferro.

Amostras	**Litologia**	**Código**	**Ferro (%)**
Bru 1	Canga	CG	61,511
Bru 2	Hematita Friável	HF	62,546
Bru 3	Hematita Friável	HF	60,906
Bru 4	Hematita Goetítica	HGO	61,282
Bru 5	Hematita Goetítica	HGO	58,759
Bru 6	Itabirito Anfibolítico	IA	53,171
Bru 7	Itabirito Anfibolítico	IA	48,515
Bru 8	Itabirito Aluminoso	IAL	57,909
Bru 9	Itabirito Aluminoso	IAL	46,690
Bru 10	Itabirito Friável	IF	56,088
Bru 11	Itabirito Friável	IF	48,602
Bru 12	Itabirito Goetítico	IGO	48,134
Bru 13	Itabirito Goetítico	IGO	37,287
Bru 14	Itabirito Magnético	IMN	59,602
Bru 15	Itabirito Magnético	IMN	52,874

4.1.3. Caracterização espectroscópica

Os testes das amostras de minério de ferro mostram características de absorção de óxidos de ferro principalmente devido à presença de minerais de hematita e goethita. A hematita é aqui caracterizada por duas bandas de absorção de óxido de ferro (III) (Fe_2O_3), a banda 670 *nm* e 900 *nm*, respectivamente. A goethita é marcada por duas bandas de absorção de óxido-hidróxido de ferro (III) $(FeO(OH))$, a banda 670 *nm* e 983 *nm*, respectivamente. As informações referentes às bandas de absorção da hematita e da goethita, e indicação das principais características de absorção decorrentes do ferro férrico podem ser observadas na Figura 4.2. Observe que a absorção apresenta um diagnóstico de hematita centrada em 869 - 880 *nm*, e goethita predominando em amostras marcadas por características em 900 - 910 *nm*.

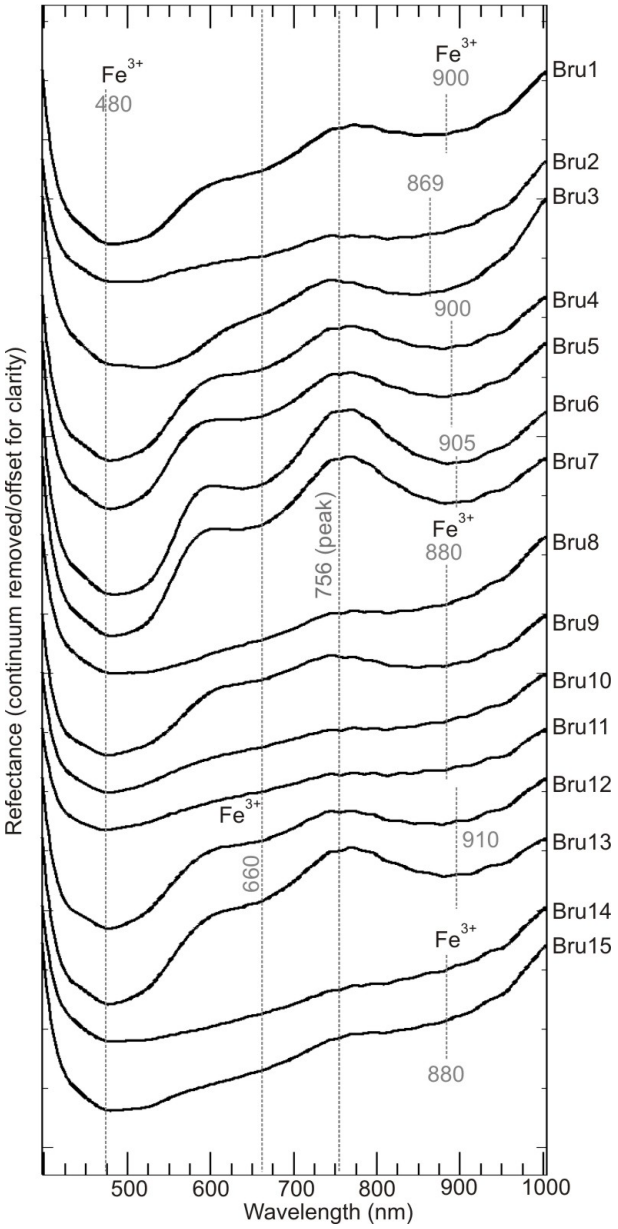


Figura 4.2: Espectros de refletância das amostras de minério de ferro.

4.1.4. Base de dados

Dos 24.000 pixeis capturados de cada uma das 15 amostras da base de dados, citados na Subseção 4.1.1, escolhemos aleatoriamente para cada amostra, através do uso de um algoritmo feito no software MATLAB®, o seguinte bloco de pixeis/amostra: 600 blocos de 40 pixeis/amostra. Reduzindo significativamente a quantidade de dados trabalhados, anteriormente 360.000 dados (15 amostras X 24.000 pixeis/amostra) para então: 9.000 dados (15 amostras X 600 blocos de 40 pixeis/amostra).

Com o intuito de reduzir os ruídos digitais de leitura de bandas espectrais de amostras, tratamos os dados utilizando a média e a mediana. Usando a média, temos uma melhor representação das características gerais da amostra, mas se houver algum valor discrepante, a média costuma tender a aumentar ou diminuir drasticamente para em torno deste valor. No entanto, quando usamos a mediana, esses valores discrepantes são suavizados, tornando a distribuição de dados mais uniforme, pois a mediana retorna a tendência central para distribuições numéricas distorcidas.

Após a realização de testes iniciais utilizando as bases de dados citadas, constatamos que a base de 9.000 dados (15 amostras X 600 blocos de 40 pixeis/amostra) que utiliza a mediana apresentou melhor desempenho ao ser manipulada pelos algoritmos de aprendizado de máquina. A base citada foi escolhida pois apresentou menor variação e melhor representação dos dados em comparação com as demais bases de dados.

Somente os teores de ferro (a serem estimados) das amostras foram considerados no modelo computacional. As entradas do modelo são as 224 bandas de cada pixel de amostra que são geradas após o processamento da imagem hiperespectral. A principal ideia é estimar os teores de ferro para cada conjunto de amostras a partir da relação de regressão encontrada pelos métodos de aprendizado de máquina entre as bandas e os teores de ferro.

4.2. Avaliação de variações de modelos

Para obtermos uma variedade de métodos de Random Forest e Multilayer Perceptron na investigação, realizamos diversos testes variando a estrutura dos métodos. Este procedimento é importante pois vai permitir visualizar e identificar possíveis ganhos ou perdas ao aumentar ou diminuir a complexidade dos parâmetros dos métodos.

A MLP descrita na Subseção 2.4.2 foi construída variando suas camadas ocultas com número de neurônios em [5,5], [10,10] e [20,20]. Para a MLP utilizamos a função da tangente hiperbólica como função de ativação, o solucionador de otimização de peso escolhido foi LBFGS que é um algoritmo de otimização na família de métodos quase-Newton que aproxima o algoritmo de Broyden-Fletcher-Goldfarb-Shanno (BFGS) usando uma quantidade limitada de memória de computador. 1.000 foi utilizado como número máximo de iterações.

A RF descrita na Subseção 2.4.3 foi construída com as seguintes variações: número de árvores em 5, 10, 20, 50, 100. A função escolhida para medir a qualidade da divisão foi o critério *mse* para o erro médio quadrático, que é igual à redução da variação como critério de seleção de recurso. E, por fim, o número mínimo de amostras necessárias para dividir um nó interno foi 2.

A base de dados de 9.000 elementos, citada na Subseção 4.1.4 foi dividida em 1/3 para validação (3.000 dados) e 2/3 (6.000 dados) para treinamento dos métodos. Sementes aleatórias foram geradas para todos os métodos durante um laço de repetição de 30 iterações, onde os métodos foram treinados e validados para a mesma base de dados.

4.3. Investigação do efeito do número de amostras

Conduzimos testes para investigarmos quais seriam os efeitos e possíveis resultados ao treinarmos o mesmo método para diferentes quantidades do número de amostras na base de dados. Para tanto, selecionamos os seguintes blocos de pixeis/amostra: 10, 20, 50, 100, 150, 300 e 600.

A justificativa e relevância desta escolha é baseada no pensamento de fornecer diferentes escolhas a partir do objetivo final, ou seja, se a escolha é performance ou melhor custo/benefício, por exemplo. Outro exemplo é fornecer escolhas para o nível de erro aceito pelo processo. Em diferentes processos de beneficiamento mineral tem-se diversos padrões aceitáveis e o percentual de erro ou acerto que se pode trabalhar para que seja adquirido o melhor resultado final.

4.4. Avaliação de variações de seleção de atributos

Muitas características hiperespectrais são redundantes devido à forte correlação entre as bandas de onda adjacentes. Portanto, a análise dos dados hiperespectrais é complexa e precisa ser simplificada, selecionando os recursos espectrais mais relevantes.

Buscando definir quais são as bandas mais significativas na estimação dos teores de ferro das amostras de minério, realizamos a avaliação da seleção de atributos no método do Random Forest de 20 árvores. Esta análise é de grande importância pois podemos verificar em quais bandas se concentram a informação relevante para os modelos de aprendizado de máquina para que possam identificar de maneira otimizada a relação existente entre as bandas da imagem hiperespectral e os teores de ferro. Isto pode significar uma potencial redução no custo do equipamento da câmera, no caso de se usar poucas bandas para realizar a estimação, assim como um potencial ganho de performance nos métodos pois seria requerido menos poder computacional para alcançar o objetivo.

5. Resultados e Discussões

Neste capítulo apresentamos os resultados alcançados pela aplicação dos métodos de aprendizado de máquina na estimação dos teores de ferro e validamos os dados de forma gráfica e estatística.

5.1. Avaliação de variações de modelos

Com o intuito de associar a massa de dados de forma estatística e obter validade e confiabilidade, computamos o coeficiente de determinação R^2, o erro médio absoluto (MAE) e o erro médio quadrático (MSE). Com essas métricas de avaliação podemos observar as relações entre a base de dados que foi separada para validação e os dados de teores de ferro estimados pelo modelo.

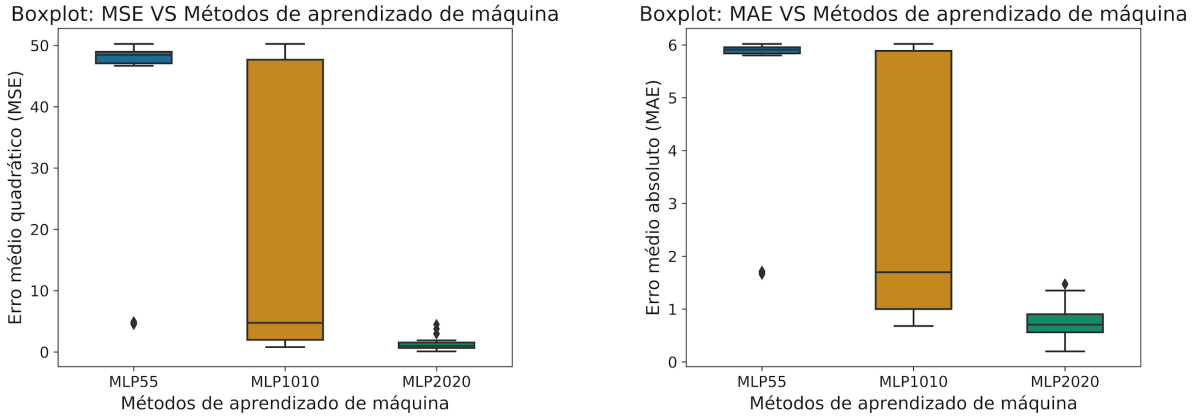
A Tabela 5.1 apresenta os resultados estatísticos de cada método para a média e desvio padrão obtidos após o treinamento das 30 rodadas. Analisando os dados da tabela, podemos verificar o desempenho superior dos RF em relação aos MLP, pois os RF possuem erros médios baixos e alto coeficiente de determinação. Além disso, podemos dar ênfase ao MSE próximo de 0 garantindo a precisão do método, pois a diferença entre os valores estimados e reais (análise química) é mínima. Os métodos de RF em suas variações propostas possuem baixo MAE e MSE, assim como possuem ótimo coeficiente de determinação (R^2). Apesar dos métodos de MLP possuírem uma boa correlação, seus resultados variam muito e seus erros são mais elevados do que o objetivado.

Tabela 5.1: Resultados das médias de 30 rodadas alcançados nos métodos de aprendizado de máquina.

Método	MSE	MAE	R^2
RF 5	0,23 $\pm$ 0,08	0,08 $\pm$ 0,01	1 $\pm$ 0,00
RF 10	0,19 $\pm$ 0,04	0,08 $\pm$ 0,01	1 $\pm$ 0,00
RF 20	0,18 $\pm$ 0,06	0,08 $\pm$ 0,01	1 $\pm$ 0,00
RF 50	0,16 $\pm$ 0,05	0,08 $\pm$ 0,01	1 $\pm$ 0,00
RF 100	0,16 $\pm$ 0,05	0,08 $\pm$ 0,01	1 $\pm$ 0,00
MLP 5,5	39,86 $\pm$ 17,89	5,08 $\pm$ 1,72	0,18 $\pm$ 0,37
MLP 10,10	18,45 $\pm$ 21,87	2,85 $\pm$ 2,25	0,62 $\pm$ 0,45
MLP 20,20	1,29 $\pm$ 1,02	0,78 $\pm$ 0,31	0,97 $\pm$ 0,02

Como as redes neurais apresentaram resultados bem abaixo dos RFs, os gráficos são mostrados separados a fim de permitir melhor interpretação visual. Perceba no eixo y das figuras a seguir que o erro é consideravelmente maior para as RNAs. A Figura 5.1 apresenta os gráficos boxplot com os respectivos valores do conjunto de repetições de cada métrica para o método do Multilayer Perceptron (MLP).

Na Figura 5.1a, que representa o erro médio quadrático (MSE), podemos verificar que o erro gira em torno de 0 e 50. O método MLP1010 apresenta maior dispersão, os dados são assimétricos positivos e possui a maior cauda de distribuição. Dentre os três, o que apresenta melhor desempenho é o MLP2020 pois possui tendência central e menor dispersão. Observando a Figura 5.1b, que representa o erro médio absoluto (MAE), podemos verificar que o erro gira em torno de 1 e 6. O método MLP1010 apresenta maior dispersão. No MLP2020 os dados possuem tendência central e houve a presença de poucos outliers. Dentre os três, o que apresenta melhor desempenho é o MLP2020. Podemos concluir que os métodos MLP55 e MPL1010 possuem desempenho insatisfatório, pois em todas as métricas apresentam variabilidade e dispersão altas. Portanto, levando em conta todas as métricas e prezando pelo melhor desempenho possível, o método MLP2020 é o que apresenta resultados melhores.



(a) Erro Médio Quadrático (MSE) - MLP.

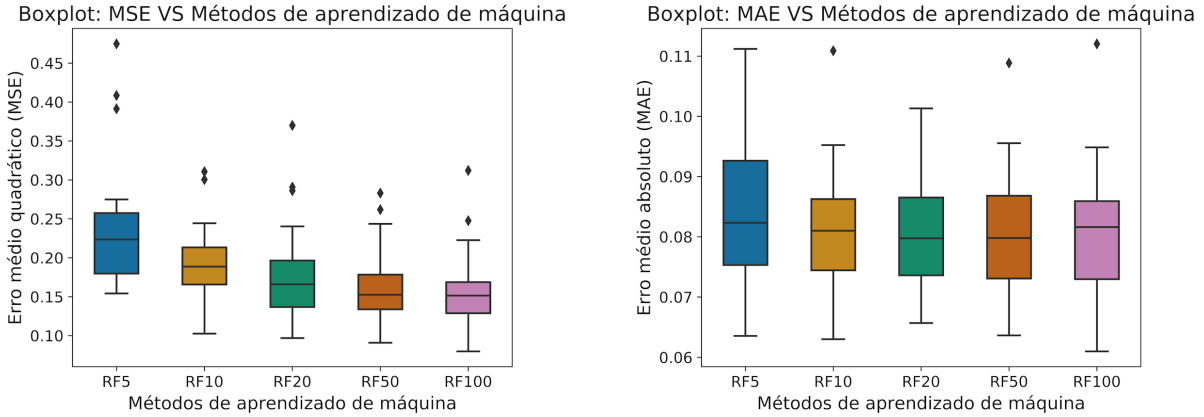
(b) Erro Médio Absoluto (MAE)- MLP.

Figura 5.1: Conjunto de gráficos boxplot das métricas estatísticas para os métodos do MLP.

A Figura 5.2 apresenta os gráficos boxplot com os respectivos valores do conjunto de repetições de cada métrica para o método RF. Analisando a Figura 5.2a, que representa o erro médio quadrático (MSE), podemos verificar que o erro gira em torno de 0,10 e 0,35. Com exceção do RF20, que possui dados bem distribuídos, tendência central e baixa dispersão, todos as variações apresentam outliers e assimetrias. Neste caso, o melhor desempenho foi alcançado pelo RF20.

Conforme a Figura 5.2b, que representa o erro médio absoluto (MAE), podemos verificar que o erro gira em torno de 0,07 e 0,11. Os métodos RF5 e RF100 apresentam outliers acima do limite máximo de detecção. O método com maior dispersão é o RF10 e o RF20 segue com o melhor desempenho entre todos.

Podemos concluir que o método RF20 possui o melhor desempenho pois em todas as métricas apresenta baixa variabilidade e dispersão. Comparando os resultados de todas as métricas estatísticas dos Random Forest com os do Multilayer Perceptron percebemos o desempenho superior dos RF. Em todos os casos, o RF possui erros menores e muito menos variabilidade, apresenta também maior centralidade dos dados, o que garante a confiabilidade de seus resultados. Assim, apenas resultados de RF serão considerados a seguir.



(a) Erro Médio Quadrático (MSE) - RF.

(b) Erro Médio Absoluto (MAE) - RF.

Figura 5.2: Conjunto de gráficos boxplot das métricas estatísticas para os métodos do RF.

Buscando definir qual o melhor conjunto de árvores para o RF, realizamos um teste estatístico sobre os resultados utilizando o software *R Version 3.5.0 Copyright © 2018 The R Foundation for Statistical Computing*. Inicialmente, avaliamos se os dados podem ser representados como distribuições normais por meio do teste de *Shapiro-Wilk normality test*. Este teste mostrou que os resultados de RF5 e RF100 não podem ser considerados distribuições normais (p-valor inferior a 0.05). Os demais podem ser considerados como distribuições normais (p-valor acima de 0.05).

A menor mediana encontrada foi utilizando o RF100, assim, comparamos todos os demais conjuntos com o RF100 buscando encontrar se algum outro conjunto é estatisticamente equivalente. Como o teste de normalidade apresentou que alguns conjuntos não podem ser considerados como destruições normais, usamos o teste não paramétrico *Wilcoxon rank sum test with continuity correction* para a avaliação de equivalência entre os conjuntos.

Os testes entre RF100, RF50 e RF20 apresentam p-valor superior a 0.05, ou seja, considerando 95% de confiança, os conjuntos podem ser considerados equivalentes. O teste entre RF20 e RF10 resultou em p-valor de 0.014, ou seja, inferior a 0,05. Considerando 95% de confiança, os conjuntos RF10 e RF20 não podem ser considerados equivalentes. Dessa forma, definimos que o melhor conjunto é o RF20. [(RF100, RF50: p-valor 0,501 > 0,05), (RF100, RF20: p-valor 0,125 > 0,05), (RF50, RF20: p-valor 0,125 > 0,05), (RF20, RF10: p-valor 0,014 < 0,05)].

Nas figuras 5.3 e 5.4, os gráficos apresentam no eixo das abcissas o número total de amostras utilizadas para a validação dos dados (3000) e no eixo das ordenadas estão os teores de ferro dos minérios. As amostras são representadas pelo código BRU de 1 a 15 e nos gráficos estão ordenadas de maneira crescente com base no teor de ferro. O conjunto de dados de cada amostra é composto por 200 pixeis/amostra, sendo que este conjunto possui o mesmo teor de ferro para todos os seus pixeis. O teor de ferro da análise química é representado pelo círculo preto e o teor de ferro estimado pelo modelo é representado pelo x azul. O círculo preto das amostras encontram-se dispostos muito próximos uns dos outros, dando a impressão de se tratar de uma linha contínua preta onde os dados estimados (x azul) se concentram.

Na Figura 5.3 está representado o resultado gráfico para a estimação utilizando o método do Multilayer Perceptron de camadas [20,20]. Neste caso, o modelo foi bastante preciso e apresentou um pouco de dispersão em algumas amostras.

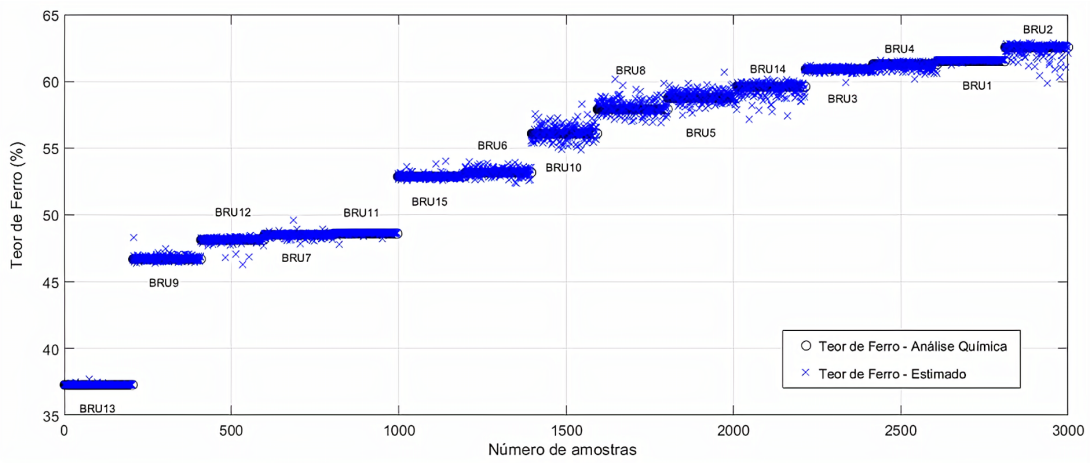


Figura 5.3: Validação do modelo do MLP de camadas [20,20] para estimação do teor de ferro.

O gráfico da Figura 5.4 apresenta o resultado para a estimação utilizando o método do Random Forest de 20 árvores. Em comparação com o método MLP, podemos notar o desempenho superior do RF ao estimar com mais precisão e menor variabilidade os teores de ferro.

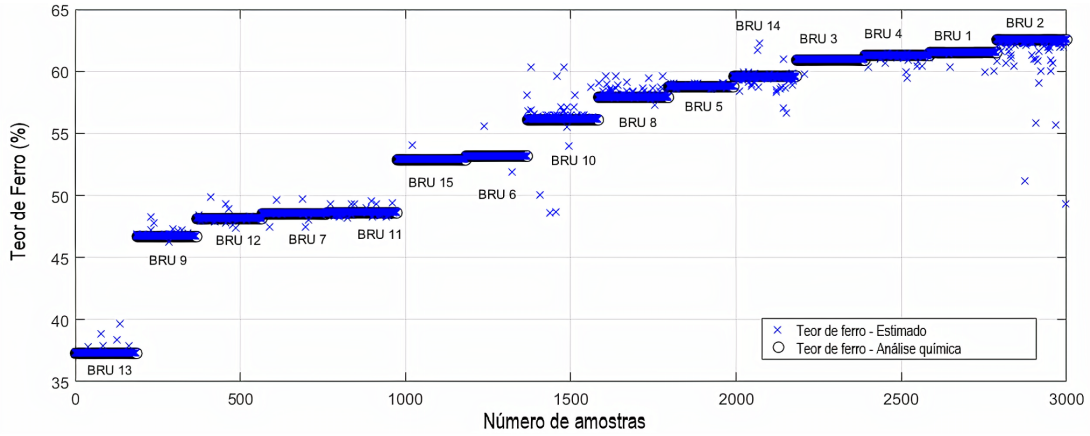


Figura 5.4: Validação do modelo do RF de 20 árvores para estimação do teor de ferro.

5.2. Investigação do efeito do número de amostras

Buscando definir qual o efeito do número de amostras em cada conjunto de árvores para o RF de 20 árvores, realizamos um teste estatístico sobre os blocos de 10, 20, 50, 100, 150, 300 e 600 pixeis/amostra. Para tanto utilizamos a métrica de avalização do desempenho do erro médio quadrático (MSE), com esta podemos mensurar a acurácia dos modelos. Podemos observar na Figura 5.5 que o erro médio quadrático fica em torno de 0 a 10. Note que todos os blocos apresentam tendência central da mediana e caudas de distribuição não muito extensas.

Começando pelo bloco de 10 pixeis/amostra, podemos verificar que os resultados são mais dispersos, porém ao aumentarmos gradativamente o número de blocos de pixeis/amostra considerados como entrada no modelo notamos uma melhoria na estabilidade e diminuição da dispersão dos dados. Reduzindo significativamente o erro inicial em torno de 4 a 8 para então próximo de zero. Portanto, dado um objetivo inicial de qual seria o maior erro aceito por um modelo, podemos trabalhar com essas variações de blocos para melhor atender a um tipo de processo ou demanda. Todos os blocos possuem resultados aceitáveis e estabilidade suficiente para performar de modo a otimizar um dado processo de estimação dos teores de ferro dos minérios de ferro.

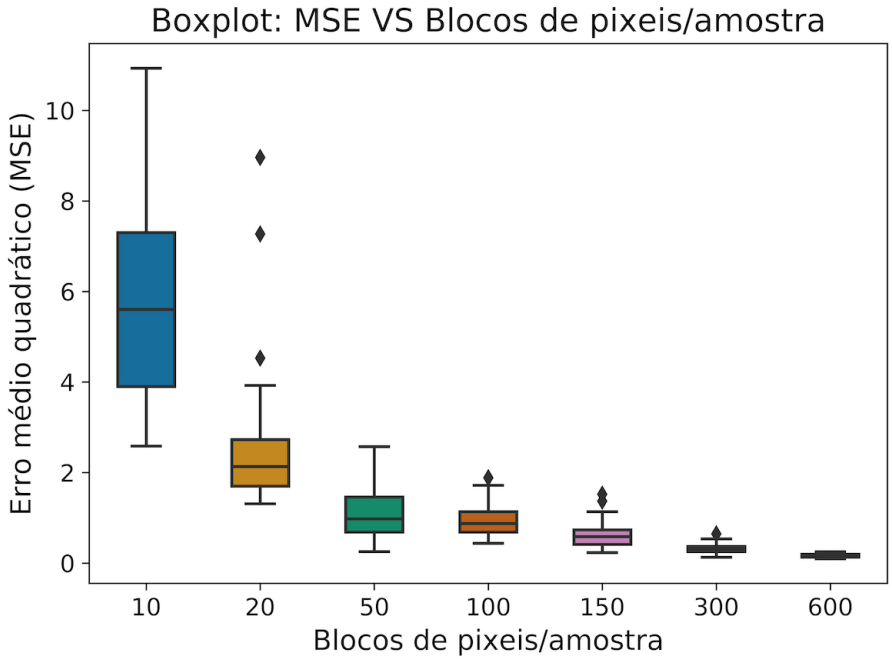


Figura 5.5: Gráfico boxplot da investigação do efeito do número de amostras na estimação dos teores de ferro dos minérios.

5.3. Avaliação de variações de seleção de atributos

Nesta etapa realizamos testes treinando o método Random Forest de 20 árvores variando o número de bandas significativas, ou seja, o número de bandas espectrais que serviram de entrada para o modelo. Para tanto, o método foi treinado utilizando as 2, 5 e 10 bandas espectrais mais significativas encontradas pelo algoritmo de seleção de bandas.

Na tabela Tabela 5.2 estão dispostas as informações referentes as 10 bandas mais relevantes na estimação dos teores de ferro dos minérios, em ordem decrescente de relevância. Nela são apresentadas a banda espectral, a frequência de onda e a relevância em porcentagem de cada banda na tarefa de estimação. Note que com as 10 bandas mais significativas foi possível obter a soma de 80,04% de relevância.

Tabela 5.2: Seleção das 10 bandas mais relevantes na estimação dos teores de ferro dos minérios.

Banda espectral	**Frequência de onda** (*nm*)	**Relevância** (%)
B658	658,40	17,52
B1000	1000,99	10,91
B930	930,78	10,53
B574	574,96	8,63
B978	978,46	7,36
B1003	1003,81	7,10
B655	655,69	7,03
B644	644,88	3,82
B642	642,18	3,60
B984	984,09	3,54

Em seguida, computamos as médias e desvio padrão de 30 rodadas do resultado do treinamento das métricas de avaliação utilizando o erro médio quadrático (MSE), o erro médio absoluto (MAE) e o coeficiente de determinação para as 224 bandas totais de entrada no modelo e para as variações com 2, 5 e 10 bandas mais significativas.

A comparação dos resultados é encontrada na Tabela 5.3. Ao analisarmos os resultados de maneira estatística, é possível notar que o melhor resultado é alcançado ao utilizar as 224 bandas como entrada no modelo, pois obtemos como resultado um MSE de 0,18, MAE de 0,08 e R^2 de 1, o que caracteriza como um desempenho ótimo.

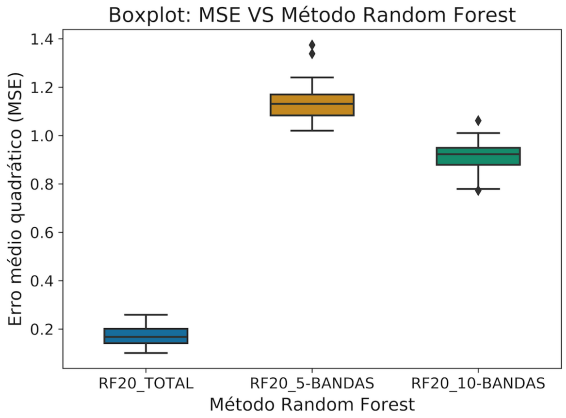
Porém, podemos notar que com exceção do uso das 2 bandas como entrada, utilizando 5 e 10 bandas também conseguimos obter resultados bastante satisfatórios e com boa precisão de

estimação dos teores. De todas as formas, também é possível observar o baixo desvio padrão de todos os métodos e métricas de avaliação.

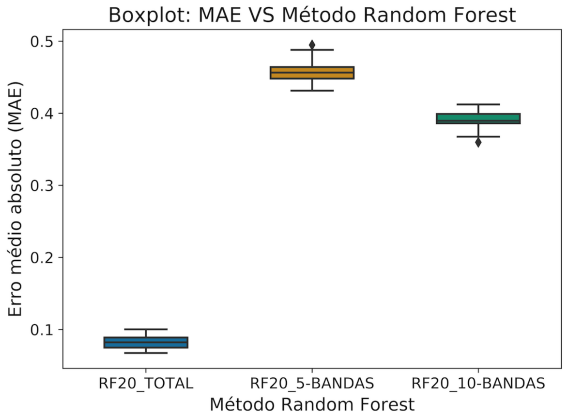
Tabela 5.3: Comparação dos resultados das médias de 30 rodadas alcançados nas variações dos de seleção de atributos para o Random Forest.

Método	MSE	MAE	R^2
RF 20 (224 bandas)	0,18 ± 0,06	0,08 ± 0,01	1 ± 0,00
RF 20 (2 bandas)	11,98 ± 0,47	2,09 ± 0,04	0,75 ± 0,01
RF 20 (5 bandas)	1,14 ± 0,08	0,46 ± 0,02	0,98 ± 0,00
RF 20 (10 bandas)	0,92 ± 0,07	0,39 ± 0,01	0,98 ± 0,00

Na Figura 5.6 é apresentado o conjunto de gráficos boxplot das métricas estatísticas para os métodos do Random Forest na seleção de atributos. Ao analisarmos os gráficos, podemos notar que os resultados alcançados são estatisticamente diferentes em ambas as métricas de avaliação.



(a) Erro Médio Quadrático (MSE) - Seleção de atributos.



(b) Erro Médio Absoluto (MAE) - Seleção de atributos.

Figura 5.6: Conjunto de gráficos boxplot das métricas estatísticas para os métodos do RF na seleção de atributos.

6. Conclusão

Os resultados demonstram que o uso do algoritmo de aprendizado de máquina Random Forest, em específico utilizando 20 árvores para estimação, apresentou melhores métricas de performance, considerando-se as análises estatísticas do coeficiente de determinação (R^2), erro médio quadrático (MSE) e percentual de acerto ao estimar os teores das amostras medidos em laboratório. Logo, ao utilizarmos o algoritmo que contém as 20 árvores, resultados satisfatórios de baixo custo e consumo computacional serão obtidos, o que se configura como vantagem competitiva e econômica. Portanto este se configura como o método mais preciso na tarefa de estimação dos teores de ferro das amostras de minérios estudadas.

Foi constado que ao variarmos o conjunto do número de amostras consideradas como entrada para o modelo, podemos notar que com quanto maior o bloco de pixeis/amostra, menor o erro médio quadrático. Por fim, foram constatadas as 10 bandas mais significativas para servir de entrada para o modelo de aprendizado de máquina do Random Forest. Em geral, o modelo conseguiu estimar com boa precisão os devidos teores de ferro medidos na análise química.

Para efeito de continuidade deste trabalho, é possível analisar a construção de outros modelos considerando diferentes métodos de aprendizado de máquina como Support Vector Machines (SVM). Além disso, é sugerido investigar o efeito que a luminosidade e umidade sobre as amostras de minério de ferro causaria nos espectros da imagem hiperespectral. Por fim, realizar a estimação dos outros teores dos elementos que compõem uma amostra de minério, como por exemplo: sílica e alumina.

Referências Bibliográficas

ABEDI, M., NOROUZI, G.-H., BAHROUDI, A. "Support vector machine for multi-classification of mineral prospectivity areas", *Computers & Geosciences*, v. 46, pp. 272–283, 2012.

ADÃO, T., HRUŠKA, J., PÁDUA, L., et al.. "Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry", *Remote Sensing*, v. 9, n. 11, pp. 1110, 2017.

ALBERS, A., MELCHIADES, F., MACHADO, R., et al.. "Um método simples de caracterização de argilominerais por difração de raios X", *Cerâmica*, v. 48, n. 305, pp. 34–37, 2002.

ALELYANI, S., TANG, J., LIU, H. "Feature selection for clustering: A review". Em: *Data Clustering*, Chapman and Hall/CRC, pp. 29–60, 2018.

ALPAYDIN, E. *Machine learning: the new AI*. MIT press, 2016.

ANTHONY, J. W., BIDEAUX, R. A., BLADH, K. W., et al.. *Handbook of mineralogy*, v. 1. Mineral Data Publ. Tucson, 1990.

AO, Y., LI, H., ZHU, L., et al.. "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling", *Journal of Petroleum Science and Engineering*, v. 174, pp. 776–789, 2019.

BARBERY, G. *Mineral liberation: measurement, simulation and practical use in mineral processing*. Québec: Éditions GB, 1991.

BIAU, G., SCORNET, E. "A random forest guided tour", *Test*, v. 25, n. 2, pp. 197–227, 2016.

BREIMAN, L. "Random forests", *Machine learning*, v. 45, n. 1, pp. 5–32, 2001.

BROWNLEE, J. "Overfitting and underfitting with machine learning algorithms", *Machine Learning Mastery*, v. 21, 2016.

CAI, J., LUO, J., WANG, S., et al.. "Feature selection in machine learning: A new perspective", *Neurocomputing*, v. 300, pp. 70–79, 2018.

CHEN, S. H., JAKEMAN, A. J., NORTON, J. P. "Artificial intelligence techniques: an introduction to their use for modelling environmental systems", *Mathematics and computers in simulation*, v. 78, n. 2-3, pp. 379–400, 2008.

COLTHUP, N. *Introduction to infrared and Raman spectroscopy*. Elsevier, 2012.

CUTLER, A., CUTLER, D. R., STEVENS, J. R. "Random forests". Em: *Ensemble machine learning*, Springer, pp. 157–175, 2012.

DE SOUZA JR, P. A. "Automation in Mössbauer spectroscopy data analysis", *Laboratory Robotics and Automation*, v. 11, n. 1, pp. 3–23, 1999.

DEDAVID, B. A., GOMES, C. I., MACHADO, G. *Microscopia eletrônica de varredura: aplicações e preparação de amostras: materiais poliméricos, metálicos e semicondutores*. EdiPUCRS, 2007.

DENG KEWANG, LI NA, Z. H. "Multilayer perceptron hyperspectral mineral classification method based on spectral absorption index". CN Patent 110031414A, Jun. 2019.

DERRICK, M. R., STULIK, D., LANDRY, J. M., et al.. *Infrared spectroscopy in conservation science*. Getty Publications, 2000.

DOMINGOS, P. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books, 2015.

DONG, X., YAN, B., GAN, F., et al.. "Progress and prospectives on engineering application of hyperspectral remote sensing for geology and mineral resources". Em: *Fifth Symposium on Novel Optoelectronic Detection Technology and Application*, v. 11023, p. 110232Y. International Society for Optics and Photonics, 2019.

DUARTE, L. D. C., JUCHEM, P. L., PULZ, G. M., et al.. "Aplicações de microscopia eletrônica de varredura (MEV) e sistema de energia dispersiva (EDS) no estudo de gemas: exemplos brasileiros", *Pesquisas em Geociências*, v. 30, n. 2, pp. 3–15, 2003.

DUNJKO, V., BRIEGEL, H. J. "Machine learning & artificial intelligence in the quantum domain: a review of recent progress", *Reports on Progress in Physics*, v. 81, n. 7, pp. 074001, 2018.

FERRARO, J. R. *Introductory raman spectroscopy*. Elsevier, 2003.

FIGUEIREDO FILHO, D. B., SILVA JÚNIOR, J. A. D. "Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r)", 2009.

GASPAR, O. D. C., OTHERS. "Microscopia e petrologia de minérios aplicadas à génese, exploração e mineralurgia dos sulfuretos maciços dos jazigos de Aljustrel e Neves-Corvo", 1995.

GEWALI, U. B., MONTEIRO, S. T., SABER, E. "Machine learning based hyperspectral image analysis: a survey", *arXiv preprint arXiv:1802.08701*, 2018.

GOMES, D. D. S. "Inteligência Artificial: conceitos e aplicaçoes", *Olhar Científico*, v. 1, n. 2, pp. 234–246, 2011.

GOWEN, A., O'DONNELL, C., CULLEN, P., et al.. "Hyperspectral imaging–an emerging process analytical tool for food quality and safety control", *Trends in food science & technology*, v. 18, n. 12, pp. 590–598, 2007.

GUPTA, H. V., KLING, H., YILMAZ, K. K., et al.. "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling", *Journal of hydrology*, v. 377, n. 1-2, pp. 80–91, 2009.

GUPTA, R. P. *Remote sensing geology*. Springer, 2017.

GY, P. *Sampling of particulate materials theory and practice*, v. 6. Elsevier, 2012.

HABOUDANE, D., MILLER, J. R., PATTEY, E., et al.. "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture", *Remote sensing of environment*, v. 90, n. 3, pp. 337–352, 2004.

HAYKIN, S. S., HAYKIN, S. S., HAYKIN, S. S., et al.. *Neural networks and learning machines*, v. 3. Pearson education Upper Saddle River, 2009.

HENLEY, K. "Ore-dressing mineralogy-a review of techniques, applications and recent developments". Em: *ICAM 81*, 1983.

HORNING, N., OTHERS. "Random Forests: An algorithm for image classification and generation of continuous fields data sets". Em: *Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan*, v. 911, 2010.

HU, P., LIU, X., CAI, Y., et al.. "Band selection of hyperspectral images using multiobjective optimization-based sparse self-representation", *IEEE Geoscience and Remote Sensing Letters*, v. 16, n. 3, pp. 452–456, 2019.

JABBAR, H., KHAN, R. Z. "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)", *Computer Science, Communication and Instrumentation Devices*, 2015.

KARR, C. *Infrared and Raman spectroscopy of lunar and terrestrial minerals*. Elsevier, 2013.

KLEIN, C., DUTROW, B. *Manual de ciência dos minerais*. Bookman Editora, 2009.

KLINGELHOEFER, G., MORRIS, R. V., BERNHARDT, B., et al.. "Athena MIMOS II Mössbauer spectrometer investigation", *Journal of Geophysical Research: Planets*, v. 108, n. E12, 2003.

KUMAR, A., MEHTA, S., VIJAYKEERTHY, D. "An Introduction to Adversarial Machine Learning". Em: *International Conference on Big Data Analytics*, pp. 293–299. Springer, 2017.

LANDGREBE, D. A. *Signal theory methods in multispectral remote sensing*, v. 29. John Wiley & Sons, 2005.

LEITE, E. P., DE SOUZA FILHO, C. R. "Artificial neural networks applied to mineral potential mapping for copper-gold mineralizations in the Carajás Mineral Province, Brazil", *Geophysical Prospecting*, v. 57, n. 6, pp. 1049–1065, 2009.

LI, J., CHENG, K., WANG, S., et al.. "Feature selection: A data perspective", *ACM Computing Surveys (CSUR)*, v. 50, n. 6, pp. 1–45, 2017.

LI JIAOJIAO, LI YUNSONG, S. L. "Hyperspectral image classification method based on spatial information enhancement and deep belief network". CN Patent 107145830A, Set. 2017.

LIU HONGCHENG, MENG SHU, Q. J. "Method for automatically identifying rock minerals based on spectral information". CN Patent 109283148A, Jan. 2019.

MAKANTASIS, K., KARANTZALOS, K., DOULAMIS, A., et al.. "Deep supervised learning for hyperspectral data classification through convolutional neural networks". Em: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 4959–4962. IEEE, 2015.

MARÔCO, J. *Análise Estatística com o SPSS Statistics.: 7ª edição*. ReportNumber, Lda, 2018.

MCCREERY, R. L. *Raman spectroscopy for chemical analysis*, v. 225. John Wiley & Sons, 2005.

MELO, V. D. F., MATTOS, J. M. S., LIMA, V. C. "Métodos de concentração de minerais 2: 1 secundários na fração argila visando sua identificação por difratometria de raios x", *Revista Brasileira de Ciência do Solo*, v. 33, n. 3, pp. 527–539, 2009.

MENZE, B. H., KELM, B. M., MASUCH, R., et al.. "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data", *BMC bioinformatics*, v. 10, n. 1, pp. 213, 2009.

MITSUTAKE, H., POPPI, R. J., BREITKREITZ, M. C. "Raman Imaging Spectroscopy: History, Fundamentals and Current Scenario of the Technique", *Journal of the Brazilian Chemical Society*, v. 30, n. 11, pp. 2243–2258, 2019.

NEUMANN, R., SCHNEIDER, C., ALCOVER NETO, A. "Caracterização tecnológica de minérios", *Tratamento de minérios*, v. 4, pp. 55–106, 2004.

NEUMANN, R., SCHNEIDER, C. L., ALCOVER NETO, A. "Parte II: Caracterização tecnológica de minérios". CETEM/MCT, 2010.

NILSSON, N. J. *Principles of artificial intelligence*. Morgan Kaufmann, 2014.

POOLE, A. B., SIMS, I. *Concrete petrography: a handbook of investigative techniques*. Crc Press, 2016.

REES, W. G., PELLIKA, P. "Principles of remote sensing", *Remote Sensing of Glaciers. London*, 2010.

RODRIGUEZ-GALIANO, V., SANCHEZ-CASTILLO, M., CHICA-OLMO, M., et al.. "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, pp. 804–818, 2015.

RODRIGUEZ-GALIANO, V. F., GHIMIRE, B., ROGAN, J., et al.. "An assessment of the effectiveness of a random forest classifier for land-cover classification", *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 67, pp. 93–104, 2012.

SANDINO, J., PEGG, G., GONZALEZ, F., et al.. "Aerial mapping of forests affected by pathogens using UAVs, hyperspectral sensors, and artificial intelligence", *Sensors*, v. 18, n. 4, pp. 944, 2018.

SANTOS, U. J. L., PESSIN, G., DA COSTA, C. A., et al.. "AgriPrediction: A proactive internet of things model to anticipate problems and improve production in agricultural crops", *Computers and electronics in agriculture*, v. 161, pp. 202–213, 2019.

SCAFUTTO, R. D. M., DE SOUZA FILHO, C. R., DE OLIVEIRA, W. J. "Hyperspectral remote sensing detection of petroleum hydrocarbons in mixtures with mineral substrates: Implications for onshore exploration and monitoring", *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 128, pp. 146–157, 2017.

SCHRÖDER, C., KLINGELHÖFER, G., MORRIS, R. V., et al.. "Field-portable Mössbauer spectroscopy on Earth, the Moon, Mars, and beyond", *Geochemistry: Exploration, Environment, Analysis*, v. 11, n. 2, pp. 129–143, 2011.

SHANMUGANATHAN, S. “Artificial neural network modelling: An introduction”. Em: *Artificial neural network modelling*, Springer, pp. 1–14, 2016.

STUART, B. “Infrared spectroscopy”, *Kirk-Othmer Encyclopedia of Chemical Technology*, 2000.

SUDHARSAN, S., HEMALATHA, R., RADHA, S. “A Survey on Hyperspectral Imaging for Mineral Exploration using Machine Learning Algorithms”. Em: *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, pp. 206–212. IEEE, 2019.

TOTH, C., JÓŹKÓW, G. “Remote sensing platforms and sensors: A survey”, *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 115, pp. 22–36, 2016.

WATT, J., BORHANI, R., KATSAGGELOS, A. *Machine learning refined: foundations, algorithms, and applications*. Cambridge University Press, 2020.

WILLMOTT, C. J., MATSUURA, K. “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”, *Climate research*, v. 30, n. 1, pp. 79–82, 2005.

WILLS, B. A., FINCH, J. *Wills' mineral processing technology: an introduction to the practical aspects of ore treatment and mineral recovery*. Butterworth-Heinemann, 2015.

ZHAO, C., GAO, Y., HE, J., et al.. “Recognition of driving postures by multiwavelet transform and multilayer perceptron classifier”, *Engineering Applications of Artificial Intelligence*, v. 25, n. 8, pp. 1677–1686, 2012.

ZHOU, Q., ZHOU, H., LI, T. “Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features”, *Knowledge-based systems*, v. 95, pp. 1–11, 2016.

ZHU, H., CHU, B., ZHANG, C., et al.. “Hyperspectral imaging for presymptomatic detection of tobacco disease with successive projections algorithm and machine-learning classifiers”, *Scientific reports*, v. 7, n. 1, pp. 1–12, 2017.

Apêndice A: Código Fonte dos Algorítimos de Aprendizado de Máquina

```
# Importação das bibliotecas necessárias para
# métricas estatísticas e métricas de aprendizado de máquina

import pandas as pd
import numpy as np
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPRegressor

# Definindo função para calcular o MAPE
def mean_absolute_percentage_error(y_true, y_pred):
    y_true, y_pred = np.array(y_true), np.array(y_pred)
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100

# Definição dos vetores para armazenar os resultados dos MSE
result_array_mse_RF5 = np.array([])
result_array_mse_RF10 = np.array([])
result_array_mse_RF20 = np.array([])
result_array_mse_RF50 = np.array([])
result_array_mse_RF100 = np.array([])
result_array_mse_MLP55 = np.array([])
result_array_mse_MLP1010 = np.array([])
result_array_mse_MLP2020 = np.array([])

# Definição dos vetores para armazenar os resultados dos MAE
result_array_mae_RF5 = np.array([])
result_array_mae_RF10 = np.array([])
result_array_mae_RF20 = np.array([])
result_array_mae_RF50 = np.array([])
result_array_mae_RF100 = np.array([])
result_array_mae_MLP55 = np.array([])
result_array_mae_MLP1010 = np.array([])
result_array_mae_MLP2020 = np.array([])

```

```
# Definição dos vetores para armazenar os resultados dos MAPE
result_array_mape_RF5 = np.array([])
result_array_mape_RF10 = np.array([])
result_array_mape_RF20 = np.array([])
result_array_mape_RF50 = np.array([])
result_array_mape_RF100 = np.array([])
result_array_mape_MLP55 = np.array([])
result_array_mape_MLP1010 = np.array([])
result_array_mape_MLP2020 = np.array([])

# Definição dos vetores para armazenar os resultados dos R2
result_array_r2_RF5 = np.array([])
result_array_r2_RF10 = np.array([])
result_array_r2_RF20 = np.array([])
result_array_r2_RF50 = np.array([])
result_array_r2_RF100 = np.array([])
result_array_r2_MLP55 = np.array([])
result_array_r2_MLP1010 = np.array([])
result_array_r2_MLP2020 = np.array([])


# Base de Dados, lendo arquivo .CSV
dataset = pd.read_csv('Data.csv')

x = dataset.iloc[:, 1:].values  #Preditores
y = dataset.iloc[:, 0].values  #A serem estimados
y = y.reshape(-1, 1)

# Laço de repetição variando a semente e treinando os métodos de IA
for i in range(1,31):

    # Dividindo a base de dados em 1/3 para validação e 2/3 para testes
    x_train, x_test, y_train, y_test = train_test_split(x, y,
    test_size = 1/3, random_state=None)

    # Definindo os parâmetros do Random Forest
    regressor_RF_5 = RandomForestRegressor(
                     n_estimators=5,
                     criterion='mse',
                     random_state=None)

    regressor_RF_10 = RandomForestRegressor(
                     n_estimators=10,
                     criterion='mse',
                     random_state=None)

    regressor_RF_20 = RandomForestRegressor(
                     n_estimators=20,
```

```
                    criterion='mse',
                    random_state=None)

    regressor_RF_50 = RandomForestRegressor(
                    n_estimators=50,
                    criterion='mse',
                    random_state=None)

    regressor_RF_100 = RandomForestRegressor(
                    n_estimators=100,
                    criterion='mse',
                    random_state=None)


    # Definindo os parâmetros do Multilayer Perceptron
    regressor_MLP_55 = MLPRegressor(
                hidden_layer_sizes=(5,5),
                activation='tanh',
                solver='lbfgs',
                max_iter=1000,
                random_state=None)

    regressor_MLP_1010 = MLPRegressor(
                hidden_layer_sizes=(10,10),
                activation='tanh',
                solver='lbfgs',
                max_iter=1000,
                random_state=None)

    regressor_MLP_2020 = MLPRegressor(
                hidden_layer_sizes=(20,20),
                activation='tanh',
                solver='lbfgs',
                max_iter=1000,
                random_state=None)

    # Realiza o ajuste do modelo com os dados de treinamento
    regressor_RF_5.fit(x_train, y_train.ravel())
    regressor_RF_10.fit(x_train, y_train.ravel())
    regressor_RF_20.fit(x_train, y_train.ravel())
    regressor_RF_50.fit(x_train, y_train.ravel())
    regressor_RF_100.fit(x_train, y_train.ravel())

    regressor_MLP_55.fit(x_train, y_train.ravel())
    regressor_MLP_1010.fit(x_train, y_train.ravel())
    regressor_MLP_2020.fit(x_train, y_train.ravel())
```

```
132     # Realiza a predição dos dados para os modelos
133     y_pred_RF5 = regressor_RF_5.predict(x_test)
134     y_pred_RF10 = regressor_RF_10.predict(x_test)
135     y_pred_RF20 = regressor_RF_20.predict(x_test)
136     y_pred_RF50 = regressor_RF_50.predict(x_test)
137     y_pred_RF100 = regressor_RF_100.predict(x_test)
138
139     y_pred_MLP55 = regressor_MLP_55.predict(x_test)
140     y_pred_MLP1010 = regressor_MLP_1010.predict(x_test)
141     y_pred_MLP2020 = regressor_MLP_2020.predict(x_test)
142
143
144     # Erro médio absoluto (MAE)
145     mae_RF5 = mean_absolute_error(y_test, y_pred_RF5)
146     mae_RF10 = mean_absolute_error(y_test, y_pred_RF10)
147     mae_RF20 = mean_absolute_error(y_test, y_pred_RF20)
148     mae_RF50 = mean_absolute_error(y_test, y_pred_RF50)
149     mae_RF100 = mean_absolute_error(y_test, y_pred_RF100)
150
151     mae_MLP55 = mean_absolute_error(y_test, y_pred_MLP55)
152     mae_MLP1010 = mean_absolute_error(y_test, y_pred_MLP1010)
153     mae_MLP2020 = mean_absolute_error(y_test, y_pred_MLP2020)
154
155     # Erro médio quadrático (MSE)
156     mse_RF5 = mean_squared_error(y_test, y_pred_RF5)
157     mse_RF10 = mean_squared_error(y_test, y_pred_RF10)
158     mse_RF20 = mean_squared_error(y_test, y_pred_RF20)
159     mse_RF50 = mean_squared_error(y_test, y_pred_RF50)
160     mse_RF100 = mean_squared_error(y_test, y_pred_RF100)
161
162     mse_MLP55 = mean_squared_error(y_test, y_pred_MLP55)
163     mse_MLP1010 = mean_squared_error(y_test, y_pred_MLP1010)
164     mse_MLP2020 = mean_squared_error(y_test, y_pred_MLP2020)
165
166     # Coeficiente de Determinação (R2)
167     r2_RF5 = r2_score(y_test, y_pred_RF5)
168     r2_RF10 = r2_score(y_test, y_pred_RF10)
169     r2_RF20 = r2_score(y_test, y_pred_RF20)
170     r2_RF50 = r2_score(y_test, y_pred_RF50)
171     r2_RF100 = r2_score(y_test, y_pred_RF100)
172
173     r2_MLP55 = r2_score(y_test, y_pred_MLP55)
174     r2_MLP1010 = r2_score(y_test, y_pred_MLP1010)
175     r2_MLP2020 = r2_score(y_test, y_pred_MLP2020)
176
177     # Erro Percentual Médio Absoluto (MAPE)
178     mape_RF5 = mean_absolute_percentage_error(y_test, y_pred_RF5)
179     mape_RF10 = mean_absolute_percentage_error(y_test, y_pred_RF10)
```

```
180     mape_RF20 = mean_absolute_percentage_error(y_test, y_pred_RF20)
181     mape_RF50 = mean_absolute_percentage_error(y_test, y_pred_RF50)
182     mape_RF100 = mean_absolute_percentage_error(y_test, y_pred_RF100)
183
184     mape_MLP55 = mean_absolute_percentage_error(y_test, y_pred_MLP55)
185     mape_MLP1010 = mean_absolute_percentage_error(y_test, y_pred_MLP1010)
186     mape_MLP2020 = mean_absolute_percentage_error(y_test, y_pred_MLP2020)
187
188
189     # Salvando resultados do MSE para exportar Excel
190     result_array_mse_RF5 = np.append(result_array_mse_RF5, mse_RF5)
191     result_array_mse_RF10 = np.append(result_array_mse_RF10, mse_RF10)
192     result_array_mse_RF20 = np.append(result_array_mse_RF20, mse_RF20)
193     result_array_mse_RF50 = np.append(result_array_mse_RF50, mse_RF50)
194     result_array_mse_RF100 = np.append(result_array_mse_RF100, mse_RF100)
195
196     result_array_mse_MLP55 = np.append(result_array_mse_MLP55, mse_MLP55)
197     result_array_mse_MLP1010 = np.append(result_array_mse_MLP1010, mse_MLP1010)
198     result_array_mse_MLP2020 = np.append(result_array_mse_MLP2020, mse_MLP2020)
199
200     # Salvando resultados do MAE para exportar Excel
201     result_array_mae_RF5 = np.append(result_array_mae_RF5, mae_RF5)
202     result_array_mae_RF10 = np.append(result_array_mae_RF10, mae_RF10)
203     result_array_mae_RF20 = np.append(result_array_mae_RF20, mae_RF20)
204     result_array_mae_RF50 = np.append(result_array_mae_RF50, mae_RF50)
205     result_array_mae_RF100 = np.append(result_array_mae_RF100, mae_RF100)
206
207     result_array_mae_MLP55 = np.append(result_array_mae_MLP55, mae_MLP55)
208     result_array_mae_MLP1010 = np.append(result_array_mae_MLP1010, mae_MLP1010)
209     result_array_mae_MLP2020 = np.append(result_array_mae_MLP2020, mae_MLP2020)
210
211     # Salvando resultados do MAPE para exportar Excel
212     result_array_mape_RF5 = np.append(result_array_mape_RF5, mape_RF5)
213     result_array_mape_RF10 = np.append(result_array_mape_RF10, mape_RF10)
214     result_array_mape_RF20 = np.append(result_array_mape_RF20, mape_RF20)
215     result_array_mape_RF50 = np.append(result_array_mape_RF50, mape_RF50)
216     result_array_mape_RF100 = np.append(result_array_mape_RF100, mape_RF100)
217
218     result_array_mape_MLP55 = np.append(result_array_mape_MLP55, mape_MLP55)
219     result_array_mape_MLP1010 = np.append(result_array_mape_MLP1010, mape_MLP1010)
220     result_array_mape_MLP2020 = np.append(result_array_mape_MLP2020, mape_MLP2020)
221
222     # Salvando resultados do R2 para exportar Excel
223     result_array_r2_RF5 = np.append(result_array_r2_RF5, r2_RF5)
224     result_array_r2_RF10 = np.append(result_array_r2_RF10, r2_RF10)
225     result_array_r2_RF20 = np.append(result_array_r2_RF20, r2_RF20)
226     result_array_r2_RF50 = np.append(result_array_r2_RF50, r2_RF50)
227     result_array_r2_RF100 = np.append(result_array_r2_RF100, r2_RF100)
```

```
228
229     result_array_r2_MLP55 = np.append(result_array_r2_MLP55, r2_MLP55)
230     result_array_r2_MLP1010 = np.append(result_array_r2_MLP1010, r2_MLP1010)
231     result_array_r2_MLP2020 = np.append(result_array_r2_MLP2020, r2_MLP2020)
232
233
234     # Cria uma pasta de trabalho Excel para cada semente
235     # Cria uma planilha dentro da pasta para cada método RF
236     # Salvando os dados do estimado e medido
237     with pd.ExcelWriter('RF_OutputSeed_{}.xlsx'.format(i)) as writer:
238
239         pd.DataFrame(y_pred_RF5, columns=['Predicted']).to_excel(writer,
240         sheet_name='RF5', index=False)
241         pd.DataFrame(y_test, columns=['Measured']).to_excel(writer,
242         sheet_name='RF5', startcol=1, index=False)
243
244         pd.DataFrame(y_pred_RF10, columns=['Predicted']).to_excel(writer,
245         sheet_name='RF10', index=False)
246         pd.DataFrame(y_test, columns=['Measured']).to_excel(writer,
247         sheet_name='RF10', startcol=1, index=False)
248
249         pd.DataFrame(y_pred_RF20, columns=['Predicted']).to_excel(writer,
250         sheet_name='RF20', index=False)
251         pd.DataFrame(y_test, columns=['Measured']).to_excel(writer,
252         sheet_name='RF20', startcol=1, index=False)
253
254         pd.DataFrame(y_pred_RF50, columns=['Predicted']).to_excel(writer,
255         sheet_name='RF50', index=False)
256         pd.DataFrame(y_test, columns=['Measured']).to_excel(writer,
257         sheet_name='RF50', startcol=1, index=False)
258
259         pd.DataFrame(y_pred_RF100, columns=['Predicted']).to_excel(writer,
260         sheet_name='RF100', index=False)
261         pd.DataFrame(y_test, columns=['Measured']).to_excel(writer,
262         sheet_name='RF100', startcol=1, index=False)
263
264
265     # Cria uma pasta de trabalho Excel para cada semente
266     # Cria uma planilha dentro da pasta para cada método MLP
267     # Salvando os dados do estimado e medido
268     with pd.ExcelWriter('MLP_OutputSeed_{}.xlsx'.format(i)) as writer:
269
270         pd.DataFrame(y_pred_MLP55, columns=['Predicted']).to_excel(writer,
271         sheet_name='MLP55', index=False)
272         pd.DataFrame(y_test, columns=['Measured']).to_excel(writer,
273         sheet_name='MLP55', startcol=1, index=False)
274
275         pd.DataFrame(y_pred_MLP1010, columns=['Predicted']).to_excel(writer,
```

```
276           ='MLP1010', index=False)
277           pd.DataFrame(y_test, columns=['Measured']).to_excel(writer,
278           sheet_name='MLP1010', startcol=1, index=False)
279
280           pd.DataFrame(y_pred_MLP2020, columns=['Predicted']).to_excel(writer,
281           sheet_name='MLP2020', index=False)
282           pd.DataFrame(y_test, columns=['Measured']).to_excel(writer,
283           sheet_name='MLP2020', startcol=1, index=False)
284
285   # Gera arquivo excel para salvar as 30 rodadas do MSE
286   with pd.ExcelWriter('MSE_Output.xlsx') as writer:
287       pd.DataFrame(result_array_mse_RF5, columns=['RF5']).to_excel(writer,
288       sheet_name='Complete', index=False)
289       pd.DataFrame(result_array_mse_RF10, columns=['RF10']).to_excel(writer,
290       sheet_name='Complete', startcol=1, index=False)
291       pd.DataFrame(result_array_mse_RF20, columns=['RF20']).to_excel(writer,
292       sheet_name='Complete', startcol=2, index=False)
293       pd.DataFrame(result_array_mse_RF50, columns=['RF50']).to_excel(writer,
294       sheet_name='Complete', startcol=3, index=False)
295       pd.DataFrame(result_array_mse_RF100, columns=['RF100']).to_excel(writer,
296       sheet_name='Complete', startcol=4, index=False)
297       pd.DataFrame(result_array_mse_MLP55, columns=['MLP55']).to_excel(writer,
298       sheet_name='Complete', startcol=5, index=False)
299       pd.DataFrame(result_array_mse_MLP1010, columns=['MLP1010']).to_excel(writer,
300       sheet_name='Complete', startcol=6, index=False)
301       pd.DataFrame(result_array_mse_MLP2020, columns=['MLP2020']).to_excel(writer,
302       sheet_name='Complete', startcol=7, index=False)
303
304   # Gera arquivo excel para salvar as 30 rodadas do MAE
305   with pd.ExcelWriter('MAE_Output.xlsx') as writer:
306       pd.DataFrame(result_array_mae_RF5, columns=['RF5']).to_excel(writer,
307       sheet_name='Complete', index=False)
308       pd.DataFrame(result_array_mae_RF10, columns=['RF10']).to_excel(writer,
309       sheet_name='Complete', startcol=1, index=False)
310       pd.DataFrame(result_array_mae_RF20, columns=['RF20']).to_excel(writer,
311       sheet_name='Complete', startcol=2, index=False)
312       pd.DataFrame(result_array_mae_RF50, columns=['RF50']).to_excel(writer,
313       sheet_name='Complete', startcol=3, index=False)
314       pd.DataFrame(result_array_mae_RF100, columns=['RF100']).to_excel(writer,
315       sheet_name='Complete', startcol=4, index=False)
316       pd.DataFrame(result_array_mae_MLP55, columns=['MLP55']).to_excel(writer,
317       sheet_name='Complete', startcol=5, index=False)
318       pd.DataFrame(result_array_mae_MLP1010, columns=['MLP1010']).to_excel(writer,
319       sheet_name='Complete', startcol=6, index=False)
320       pd.DataFrame(result_array_mae_MLP2020, columns=['MLP2020']).to_excel(writer,
321       sheet_name='Complete', startcol=7, index=False)
322
323
```

```
# Gera arquivo excel para salvar as 30 rodadas do MAPE
with pd.ExcelWriter('MAPE_Output.xlsx') as writer:
    pd.DataFrame(result_array_mape_RF5, columns=['RF5']).to_excel(writer,
    sheet_name='Complete', index=False)
    pd.DataFrame(result_array_mape_RF10, columns=['RF10']).to_excel(writer,
    sheet_name='Complete', startcol=1, index=False)
    pd.DataFrame(result_array_mape_RF20, columns=['RF20']).to_excel(writer,
    sheet_name='Complete', startcol=2, index=False)
    pd.DataFrame(result_array_mape_RF50, columns=['RF50']).to_excel(writer,
    sheet_name='Complete', startcol=3, index=False)
    pd.DataFrame(result_array_mape_RF100, columns=['RF100']).to_excel(writer,
    sheet_name='Complete', startcol=4, index=False)
    pd.DataFrame(result_array_mape_MLP55, columns=['MLP55']).to_excel(writer,
    sheet_name='Complete', startcol=5, index=False)
    pd.DataFrame(result_array_mape_MLP1010, columns=['MLP1010']).to_excel(writer,
    sheet_name='Complete', startcol=6, index=False)
    pd.DataFrame(result_array_mape_MLP2020, columns=['MLP2020']).to_excel(writer,
    sheet_name='Complete', startcol=7, index=False)

# Gera arquivo excel para salvar as 30 rodadas do R2
with pd.ExcelWriter('R2_Output.xlsx') as writer:
    pd.DataFrame(result_array_r2_RF5, columns=['RF5']).to_excel(writer,
    sheet_name='Complete', index=False)
    pd.DataFrame(result_array_r2_RF10, columns=['RF10']).to_excel(writer,
    sheet_name='Complete', startcol=1, index=False)
    pd.DataFrame(result_array_r2_RF20, columns=['RF20']).to_excel(writer,
    sheet_name='Complete', startcol=2, index=False)
    pd.DataFrame(result_array_r2_RF50, columns=['RF50']).to_excel(writer,
    sheet_name='Complete', startcol=3, index=False)
    pd.DataFrame(result_array_r2_RF100, columns=['RF100']).to_excel(writer,
    sheet_name='Complete', startcol=4, index=False)
    pd.DataFrame(result_array_r2_MLP55, columns=['MLP55']).to_excel(writer,
    sheet_name='Complete', startcol=5, index=False)
    pd.DataFrame(result_array_r2_MLP1010, columns=['MLP1010']).to_excel(writer,
    sheet_name='Complete', startcol=6, index=False)
    pd.DataFrame(result_array_r2_MLP2020, columns=['MLP2020']).to_excel(writer,
    sheet_name='Complete', startcol=7, index=False)
```

Apêndice B: Código Fonte do Algorítimo de Seleção de Bandas

```
#Importação das bibliotecas necessárias para
# métricas estatísticas e métricas de aprendizado de máquina

import pandas as pd
import numpy as np
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import SelectFromModel

# Definindo função para calcular o MAPE
def mean_absolute_percentage_error(y_true, y_pred):
    y_true, y_pred = np.array(y_true), np.array(y_pred)
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100

# Definição do vetor para armazenar os resultados dos MSE
result_array_mse_RF20 = np.array([])

# Definição do vetor para armazenar os resultados dos MAE
result_array_mae_RF20 = np.array([])

# Definição do vetor para armazenar os resultados dos MAPE
result_array_mape_RF20 = np.array([])

# Definição do vetor para armazenar os resultados dos R2
result_array_r2_RF20 = np.array([])

# Base de Dados, lendo arquivo .CSV
dataset = pd.read_csv('Data.csv')

x = dataset.iloc[:, 1:].values  #Preditores
y = dataset.iloc[:, 0].values  #A serem estimados
y = y.reshape(-1, 1)


```

```
# Pegando nome dos atributos (bandas)
feat_labels = list(dataset)

# Dividindo a base de dados em 1/3 para validação e 2/3 para testes
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 1/3, random_state=None)

# Definindo os parâmetros do Random Forest
regressor_RF_20 = RandomForestRegressor(
                n_estimators=20,
                criterion='mse',
                random_state=None)


# Realiza o ajuste do modelo com os dados de treinamento
regressor_RF_20.fit(x_train, y_train.ravel())

a_list = []
# Pega os nomes e significancia de cada atributo
for feature in zip(feat_labels, regressor_RF_20.feature_importances_):
    a_list.append(feature)

# Cria um objeto seletor que vai usar o método do RF para
# selecionar os atributos com maior significancia
sfm = SelectFromModel(regressor_RF_20, threshold=0.05)

# Treina o seletor
sfm.fit(x_train, y_train.ravel())

# Exibe os atributos selecionados com maior importancia
for feature_list_index in sfm.get_support(indices=True):
    print (a_list[feature_list_index])


# Transforma os dados para criar um novo dataset contendo somente os atributos mais importantes
X_important_train = sfm.transform(x_train)
X_important_test = sfm.transform(x_test)

# Definindo os parâmetros do Random Forest
regressor_RF_20_important = RandomForestRegressor(
                n_estimators=20,
                criterion='mse',
                random_state=None)

regressor_RF_20_important.fit(X_important_train, y_train.ravel())

# Realiza a predição dos dados para os modelos
y_pred_RF20 = regressor_RF_20.predict(x_test)
y_pred_RF20_important = regressor_RF_20_important.predict(X_important_test)
```

```
84
85  # Erro médio absoluto (MAE)
86  mae_RF20 = mean_absolute_error(y_test, y_pred_RF20)
87  mae_RF20_important = mean_absolute_error(y_test, y_pred_RF20_important)
88
89  # Erro médio quadrático (MSE)
90  mse_RF20 = mean_squared_error(y_test, y_pred_RF20)
91  mse_RF20_important = mean_squared_error(y_test, y_pred_RF20_important)
92
93  # Coeficiente de Determinação (R2)
94  r2_RF20 = r2_score(y_test, y_pred_RF20)
95  r2_RF20_important = r2_score(y_test, y_pred_RF20_important)
96
97  # Erro Percentual Médio Absoluto (MAPE)
98  mape_RF20 = mean_absolute_percentage_error(y_test, y_pred_RF20)
99  mape_RF20_important = mean_absolute_percentage_error(y_test, y_pred_RF20_important)
```