Eduardo Sartori Vieira Carvalho Leme

APLICAÇÃO DE MACHINE LEARNING NA IDENTIFICAÇÃO DE HORIZONTES DO REGOLITO NO DEPÓSITO DE NÍQUEL LATERÍTICO DO RIO DOS BOIS, GOIÁS

Trabalho Final de Curso

(Geologia)

UFRJ
Rio de Janeiro
2023

Eduardo Sartori Vieira Carvalho Leme

APLICAÇÃO DE MACHINE LEARNING NA IDENTIFICAÇÃO DE HORIZONTES DO REGOLITO DO DEPÓSITO DE NÍQUEL LATERÍTICO DO RIO DOS BOIS, GOIÁS

Trabalho de Conclusão de Curso de Graduação em Geologia do Instituto de Geociências, Universidade Federal do Rio de Janeiro – UFRJ, apresentado como requisito necessário para a obtenção do grau de Bacharel em Geologia.

Orientador(es):

Claudio Gerheim Porto – UFRJ

Rio de Janeiro
JANEIRO/2023

Eduardo Sartori Vieira Carvalho Leme

Aplicação de Machine Learning na identificação de horizontes do regolito no depósito de níquel laterítico do Rio dos Bois, Goiás / Eduardo Sartori Vieira Carvalho Leme – Rio de Janeiro: UFRJ / IGeo, 2023.

73 f.

Trabalho Final de Curso (Geologia) – Universidade Federal do Rio de Janeiro, Instituto de Geociências, Departamento de Geologia, 2023.

Orientador(es): Claudio Gerheim Porto

1. Geologia. 2. Exploração Mineral 3. Geoestatística – Trabalho de Conclusão de Curso. I. Claudio Gerheim Porto. II. Universidade Federal do Rio de Janeiro, Instituto de Geociências, Departamento de Geologia. III. Uso do algoritmo Random Forest na previsão e correção de litologias para dados geoquímicos do depósito Rio dos Bois em Iporá – Go.

Eduardo Sartori Vieira Carvalho Leme

APLICAÇÃO DE MACHINE LEARNING NA IDENTIFICAÇÃO DE HORIZONTES DO REGOLITO NO DEPÓSITO DE NÍQUEL LATERÍTICO DO RIO DOS BOIS, GOIÁS

Trabalho de Conclusão de Curso de Graduação em Geologia do Instituto de Geociências, Universidade Federal do Rio de Janeiro – UFRJ, apresentado como requisito necessário para a obtenção do grau de Bacharel em Geologia.

Orientador(es):

Claudio Gerheim Porto – UFRJ

Aprovada em: 27/01/2023

Por:

_____

Orientador: Dr. Claudio Gerheim Porto, UFRJ

_____

Luis Paulo Braga, UFRJ

_____

José Carlos Sícole Seoane, UFRJ

UFRJ
Rio de Janeiro
2023

Agradecimentos

Agradeço a todos que fizeram parte desse capítulo da minha vida e do meu processo educacional, dedico este trabalho à vocês. Primeiramente aos meus pais Andréa e Ercilio que me deram tudo para que eu pudesse crescer com saúde, educação, amor e carinho.

Aos meus amigos de infância, do skate e da vida por me ajudarem a crescer e aprender na vida com suas companhias.

Ao meu orientador Claudio Porto que me acolheu, ensinou e proporcionou que este trabalho se tornasse realidade. Aos professores Cainho e Renata por me proporcionarem muitos ensinamentos nas monitorias. A todos professores do departamento de geologia da UFRJ, aos motoristas do IGeo e todos funcionarios que fizeram parte da minha formação.

Agradeço aos meus amigos do D.A. Joel Valença : Bernardo, Vinícius, Gil, PV, Julinha, Bianca, Luan, Locatelli, Mogli, Thauan, Rapha e muitos outros que fizeram parte das tardes no fundão e das inúmeras viagens de campo que fizemos nos divertindo e aprendendo.

Aos colegas do curso de graduação que dividiram os dias no fundão comigo. Aos colegas do BCMT que começaram junto de mim nesta jornada. Agradeço ao meu colega do laboratório de exploração mineral João Casado por contribuir com este trabalho me ajudando e ensinando para que o mesmo fosse concluido.

Agradeço pelas oportunidades que a UFRJ me ofereceu nesses 7 anos de graduação que foram essenciais para meu desenvolvimento. Sempre serei grato por ter tido a oportunidade de estudar ciências numa instituição de alto nível.

Resumo

LEME, Eduardo Sartori Vieira Carvalho. **Aplicação de Machine Learning na identificação de horizontes do regolito no depósito de níquel laterítico do Rio dos Bois, Goiás**. Rio de Janeiro, Ano. 2023, 73 f. Trabalho Final de Curso – Departamento de Geologia, Instituto de Geociências, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2023.

O depósito Rio dos Bois localizado ao norte do município de Iporá está inserido na Província Alcalina do Sul de Goiás e é descrito como um corpo alcalino ultramáfico que desenvolveu um perfil laterítico que hospeda mineralizações dos commodities Níquel e Cobalto. O corpo é uma intrusão de cerca de 70 Ma que corta as rochas pré cambrianas da região com um formato oval e é composto principalmente de dunitos serpentinizados. O presente trabalho tem como objetivo aplicar técnica de machine learning para prever e corrigir a classificação litológica de testemunhos de 737 furos de sondagem e com isso auxiliar a modelagem do depósito. Para tanto foi utilizada a base de dados geoquímicos do depósito Rio dos Bois concedida pela Teck Cominco Ltd. . O método emprega o algoritmo Random Forest utilizando o software Orange Data Mining para prever as classes litológicas usando um conjunto de dados de treino para ensina-lo sobre a distribuição geoquímica das classes. O treinamento foi feito selecionando as amostras com composição geoquímica mais representativa de cada classe litológica por meio da análise de boxplots da distribuição geoquímica dos principais metais. A previsão pode ser visualizada pelos gráficos de frequência, boxplots e RadViz que revelam diferentes zonas geoquímicas para cada classe podendo assim, corrigir a base de dados para melhor interpretação do depósito.

Palavras-chave: Níquel Laterítico; Cobalto Laterítico; Machine Learning; Random Forest; Geoquímica; Depósito Rio dos Bois.

Abstract

LEME, Eduardo Sartori Vieira Carvalho**. Aplicação de Machine Learning na identificação de horizontes do regolito no depósito de níquel laterítico do Rio dos Bois, Goiás**. [*Application of Machine Learning in the identification of regolith horizons in the lateritic nickel deposit of Rio dos Bois, Goiás*] Rio de Janeiro, Ano. 2023, 73 f. Trabalho Final de Curso –Departamento de Geologia, Instituto de Geociências, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2023.

The Rio dos Bois deposit located north of the city of Iporá is inserted in the Alkaline Province of Southern Goiás and described as an ultramafic alkaline body upon which a full lateritic profile has developed and hosts mineralizations of the commodities Nickel and Cobalt. The body is an intrusion of around 70 Ma that cuts the Precambrian rocks of the region with an oval shape and is mainly composed of serpentinized dunites. The present work aims to use machine learning to predict and correct the classification of core samples form 737 drillholes using the geochemical database of the Rio dos Bois deposit granted by Teck Cominco Ltd. and assist in modeling of the deposit. The method employs the use of the Random Forest algorithm using Orange Data Mining software to predict the lithological classes using a training dataset to teach the algorithm about the geochemical distribution of the classes. The training was done by selecting most representative samples of each lithological class. This was done through the analysis of boxplots of the geochemical distribution of the main metals that compose them. The prediction can be visualized by frequency graphs, boxplots and RadViz that reveal different geochemical zones for each class, thus being able to correct the database for better interpretation of the deposit.

Key-Words: Lateritic Nickel; Lateritic Cobalt; Machine Learning; Random Forest; Geochemistry; Rio dos Bois Deposit.

Índice de tabelas

Índice de figuras

Sumário

1 INTRODUÇÃO

O presente trabalho tem como objetivo criar um modelo preditivo de machine learning usando o algoritmo *Random Forest* através do software de código aberto *Orange Data Mining* para dados geoquímicos de poço e de descrição litológica de campo do depósito de níquel laterítico Rio dos Bois localizado a norte da cidade de Iporá na porção sudoeste do estado de Goiás, Centro-Oeste do Brasil (Figura 1) a cerca de 190 quilômetros de Goiânia. Esse modelo auxiliará a análise da base de dados a fim de corrigir possíveis classificações litológicas que não correspondem a seus grupos geoquímicos, utilizando-se de estatística e machine learning para fazer tais previsões de classe.

A relevância econômica do trabalho encontra-se justamente nas etapas exploratórias dos depósitos lateríticos de Ni e Co onde os trabalhos exploratórios visam a construção de um modelo geológico a partir da correta identificação dos horizontes do regolito. Isso vai permitir que as estimativas de recursos a partir da base de dados seja realizada com maior eficiência e acuidade.

Considerando um crescimento global das demandas para matéria-prima, da complexidade das estruturas geológicas dos depósitos minerais, e da diminuição do teor dos minérios, são necessárias informações mineralógicas abundantes e de qualidade (Jooshaki et al., 2021). O crescente uso de computadores potentes e métodos de machine learning tem sido bem explorado pela industria mineral pela grande quantidade de dados disponiveis e de tarefas para serem executadas. Assim, o emprego de máquinas para realização de tarefa complexas a partir de extensas bases de dados compreendendo inumeras variáveis, são a cada dia mais necessárias.

Figura 1- Mapa de localização do município de Iporá onde está o depósito Rio dos Bois.

2 CONTEXTO GEOLÓGICO

2.1 Geologia regional

A área de estudo pertence tectonicamente a parte central da Província Tocantins e mais especificamente no Arco Magmático de Goiás (Figura 2) e geologicamente a Província Alcalina do Sul de Goiás (Figura 3). Situa-se a sudeste do Craton Amazonico, a oeste do Craton São Francisco e a norte da Bacia do Paraná (Pimentel et al., 1999), onde dominam rochas alcalinas ultramáficas de idade Mesozóica.

Estas rochas são intrusivas no terreno Neoproterozoico que compreendem rochas do Complexo granitóide-gnáissicos, granitos pós tectônicos Rio Caiapó-Iporá, rochas metavulcano-sedimentares de Bom Jardim de Goiás, Arenópolis-Piranhas e Iporá-Amorinópolis e nas rochas Fanerozóicas sedimentares da Bacia do Paraná da Formação Furnas e Ponta Grossa.

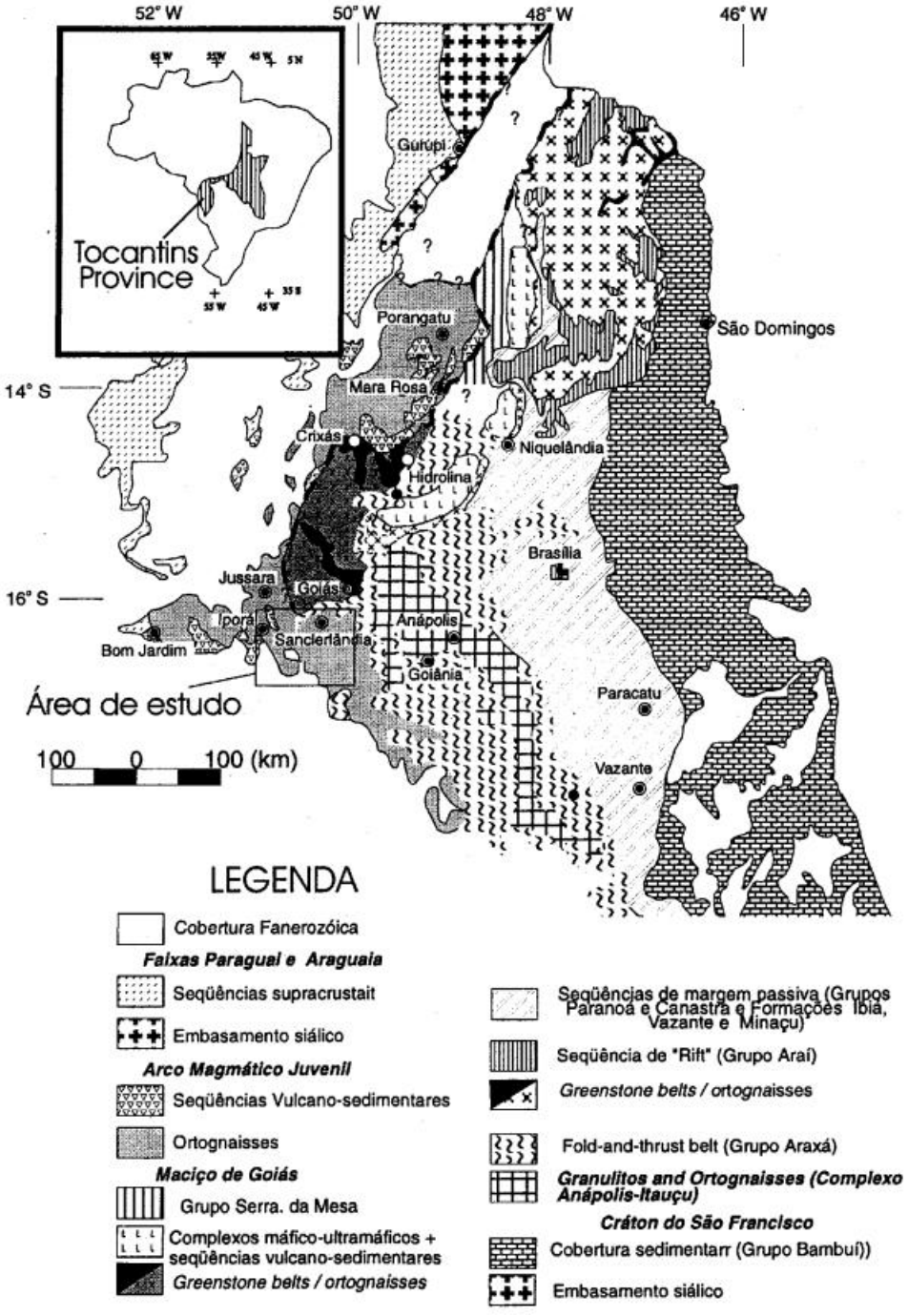

Figura 2: Mapa regional da Província Tocantins. Fonte: Pimentel et al., 1999.

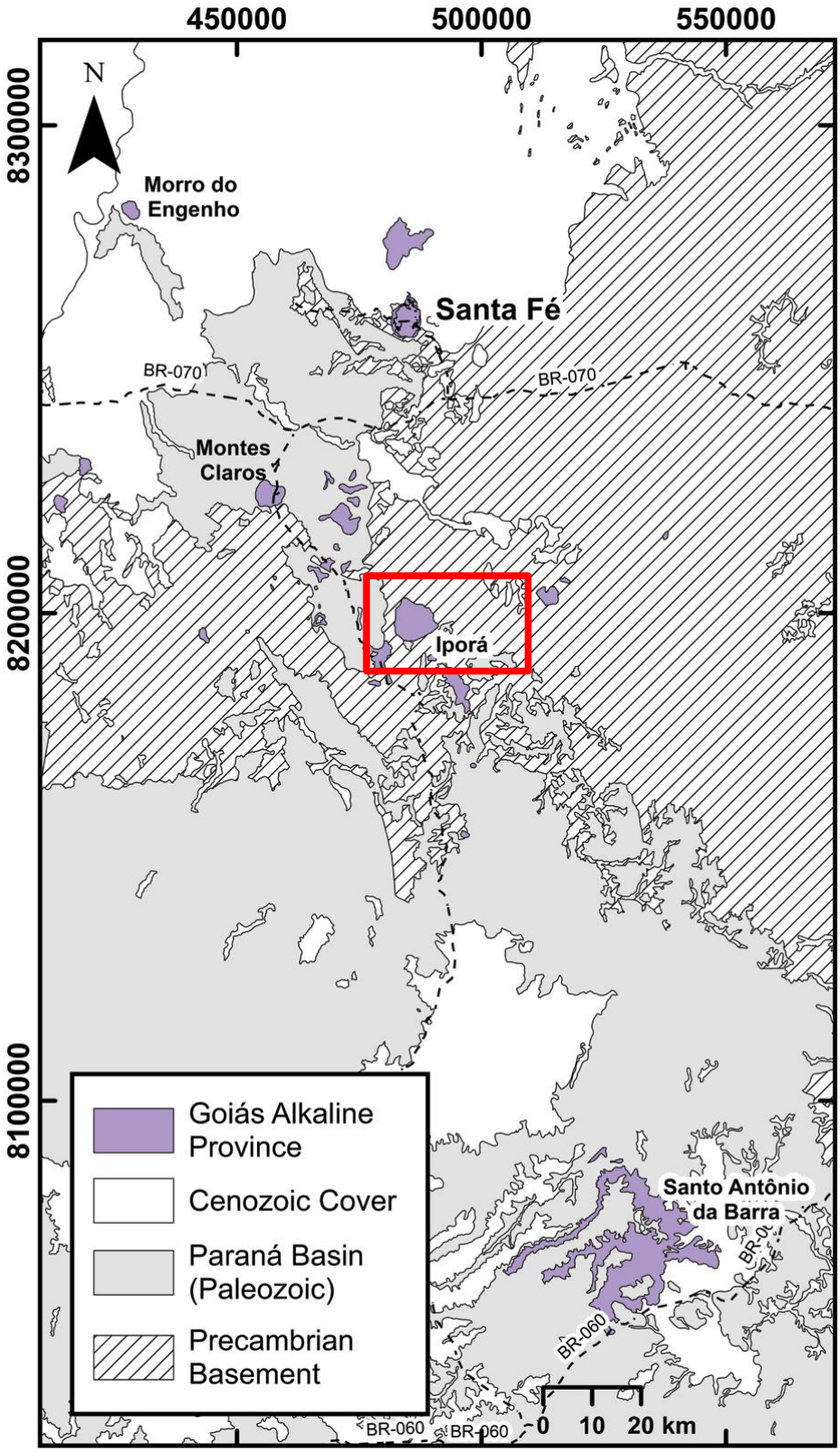

Figura 3: Mapa geológico do sul de Goiás mostrando a localização da Província Alcalina de Goiás. Retirado de Putzulo et al 2020.

2.2 Geologia do depósito

O corpo intrusivo que constitui o substrato da mineralização tem dimensões de 5km de diâmetro e 20km² de área, idade aproximada de 70 milhões de anos e é composto de diversas rochas ultramáficas intrusivas de caráter alcalino onde no centro predominam dunitos e nas bordas peridotitos, piroxênitos, gabros e sienitos. Essas rochas são cobertas por um regolito laterítico onde o mineral-minério de Ni constitui argilas Niquelíferas de oxi-hidróxidos de Fe. O mapa geológico da área de estudo encontra-se na Figura 4.

Os dunitos e peridotitos são rochas ultramelanocráticas, isotrópicas que em grande maioria se encontram serpentinizados com alto grau de silicificação e com mineralogia composta principalmente de olivinas forsteríticas com graus variados de serpentinização, piroxênios e opacos (Pena et al., 1975). A área de estudo apresenta quatro morros principais que propiciam a lixiviação dos metais para áreas mais planas ao redor (Pena et al., 1975). Em casos como Santa Fé, de acordo com Barbour (1976), as cristas dos morros são altamente silicificadas e ajudam a proteger o regolito dos agentes intempéricos atuais propiciando assim uma acumulação do Ni.

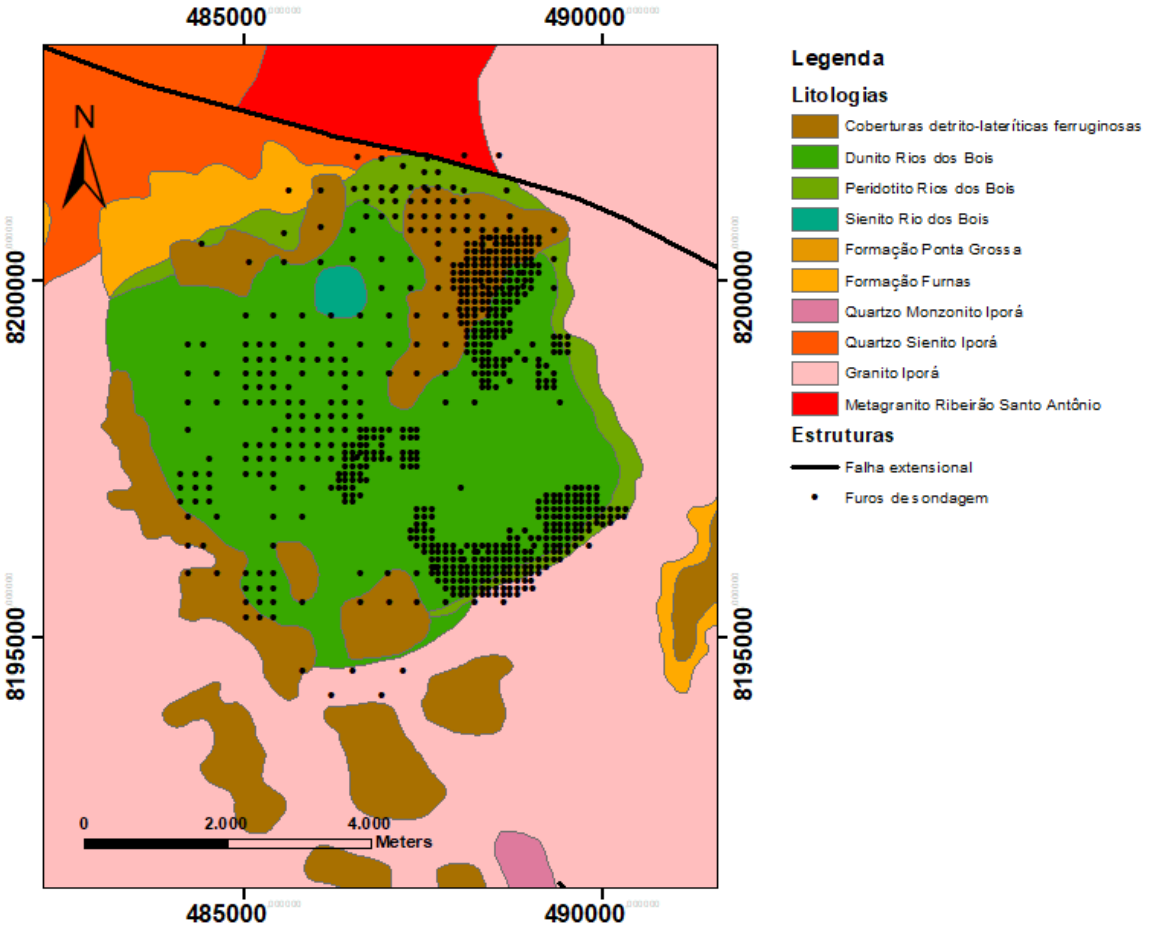

Figura 4: Mapa geológico da área de estudo com furos de sondagem. Feito sobre folha 1:100000 da CPRM.

2.3 Evolução do perfil laterítico

Um perfil laterítico niquelífero se desenvolve através do intemperismo de rochas ultramáficas niquelíferas onde o minério é resultado de dissolução e precipitação de alguns minerais através da infiltração de água meteórica mais ou menos ácida. O principal mineral portador de níquel no protólito são as olivinas magmáticas ricas em Mg (fosterita) e de seus produtos de alteração como serpentina e clorita (Golithly et al., 2010).

Os perfis se formam pela intensa atuação da água por meio da lixiviação dos minerais da rocha, carregando e remobilizando metais e concentrando outros no resíduo. A sílica e o magnésio são os principais a serem lixiviados do perfil, em contrapartida, o ferro, alumínio, cromo e titânio por serem menos móveis se concentram no resíduo no topo do perfil.

No primeiro estágio se desenvolve o saprólito pela alteração intempérica da rocha ultramáfica gerando minerais argilosos como os do grupo da garnierita, preservando estruturas primárias da rocha e formando um horizonte intempérico cujo contato com o protólito pode ser bem irregular seguindo descontinuidades dadas por fraturas e falhas onde blocos de rocha sã encontram-se imersos na matriz de material alterado (Gleeson et al. 2003). No estágio seguinte se desenvolve o horizonte oxidado onde o magnésio é inteiramente lixiviado assim como parte da sílica formando óxidos e hidróxidos de ferro e obliterando as estruturas primárias. No estágio final forma-se uma dura camada de concreção ferruginosa ou crosta laterítica que protege as camadas inferiores da erosão e concentra minerais residuais do perfil como hidróxidos de ferro. Na Figura 5 vemos um esquema de perfil laterítico niquelífero desenvolvido em rocha ultramáfica em clima tropical e suas respectivas composições químicas.

Estes perfis são controlados por diversas características que juntas formam um ambiente propício para o desenvolvimento e preservação destes depósitos. Dentre essas características se destacam a composição do protólito, clima, relevo, tectônica regional, taxa de intemperismo e nível do lençol freático.

2.3.1 Clima e relevo

Como mencionado acima, o clima e o relevo desempenham juntos um papel importante para o desenvolvimento e preservação destes depósitos. O relevo está relacionado com o nível freático e drenagem que influenciam na concentração de níquel em horizontes basais do perfil através da lixiviação de elementos móveis.

A tectônica pode ser relacionada a soerguimentos de platôs e terraços que ajudam no rebaixamento do lençol freático e consequentemente no aumento da taxa de infiltração além de promover um controle estrutural através de falhas e zonas de cisalhamento.

O clima desempenha um papel fundamental na formação destes regolitos que em geral são mais bem desenvolvidos sob a influência de climas tropicais (Indonesia e Colombia) e de savana (Nova Caledônia e Centro-Oeste do Brasil) que propiciam alta pluviosidade, umidade e atividade biogênica resultando num ambiente exclusivo para o intemperismo químico (Gleeson et al. 2003).

2.3.2 Tipos de minério

Os depósitos de níquel laterítico se caracterizam pela mineralogia dos principais minérios encontrados e se dividem em tipos como: silicato hidratado, oxidado e argila silicática (Gleeson et al., 2003 e Brand et al., 1998).

Nos depósitos de silicato hidratado o minério se hospeda na zona do saprólito tomando forma de silicatos hidratados de magnésio e níquel como garnierita onde zonas de veios e de box-works se desenvolvem através de concentração supergênica geralmente como resultado da lixiviação das fases minerais de hidróxidos de ferro do horizonte oxidado.

Fe	MgO	Ni	Co
>50%	<0.5%	<0.6%	<0.1%
40-50%	0.5-5%	0.8-1.5%	0.1-0.2%
10-25%	15-35%	1.5-3%	0.02-0.1%
5%	35-45%	0.3%	0.01%

Figura 5- Perfil laterítico esquemático desenvolvido em rocha ultrmáfica em clima tropical, mostrando a composição química em wt%. Retirado de Elias, M. 2002.

Nos depósitos dominados por óxidos ou limonitizados o minério se concentra nas fases de oxi-hidróxidos de ferro em geral na goetita e em algumas ocorrências nos óxidos de manganês que também podem conter cobalto associado às suas estruturas.

Para os depósitos de argila silicática a sílica é parcialmente removida pela água restando apenas combinações de Ni, Fe e Al que formam argilas ricas em níquel como a nontronita. Esse tipo de minério é comum na porção superior do regolito.

O depósito Rio dos Bois apresenta dois tipos de minério: silicatado e oxidado. A Figura 6 mostra os três tipos mencionados acima mostrando e a relação de espessura dos principais horizontes de cada tipo.

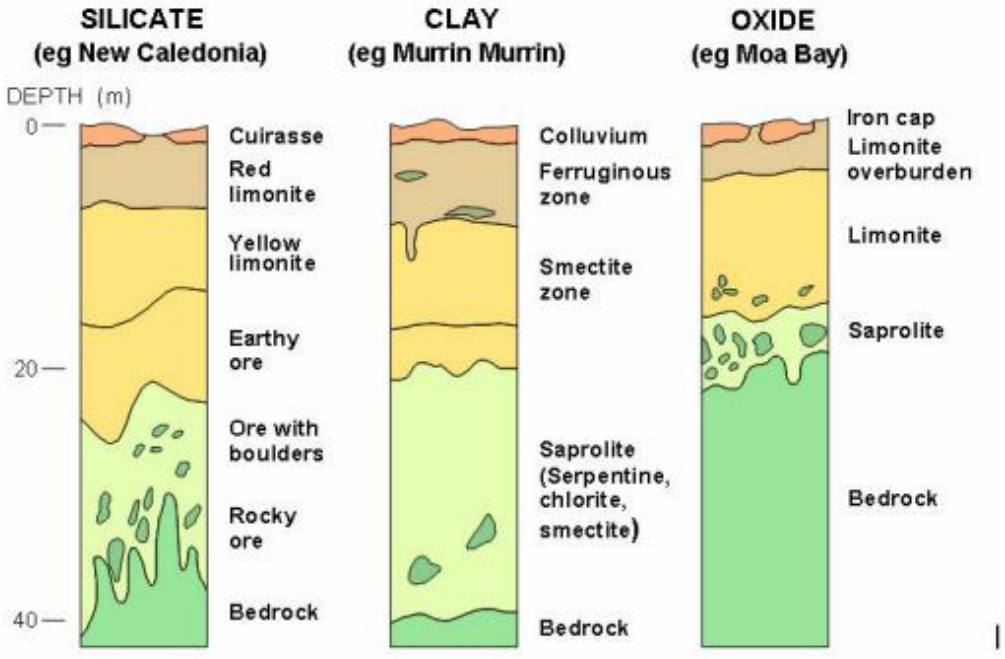

Figura 6: Comparação esquemática dos principais tipos de perfis lateríticos. Retirado de Elias, M. 2002.

3 METODOLOGIA

Para a realização do trabalho foi feita uma revisão bibliográfica de artigos envolvendo o uso do algoritmo *Random Fores*t e de outros métodos de machine learning assim como estudos feitos na região. Para o tratamento da base de dados foi utilizado o software *LibreOffice Calc versão 7.3.6.2*. e para os métodos estatísticos e de machine learning foi utilizado o software *Orange Data Mining versão 3.31.1*.

A base de dados do depósito Rio dos Bois foi concedida pela Teck Cominco Ltd. Os dados foram processados nos softwares citados acima para serem introduzidos no modelo. O modelo preditivo visa identificar a classe do dado intervalo geoquímico utilizando um conjunto de dados de treino para ensinar o algoritmo a prever a classe das amostras da base de dados.

O Orange Data Mining é um software grátisde código aberto usado para *machine learning* e visualização de dados. Sua interface permite uma simples integração das diferentes ferramentas usadas nas etapas do trabalho como, por exemplo, gráficos boxplot e ferramentas de concatenação de colunas. As ferramentas compreendem diferentes métodos estatísticos e de machine learning além de visualização e tratamento de dados.

O uso das ferramentas de visualização objetiva mostrar através dos gráficos multi-elementares e mapas altimétricos o desempenho do algoritmo para classificar as amostras.

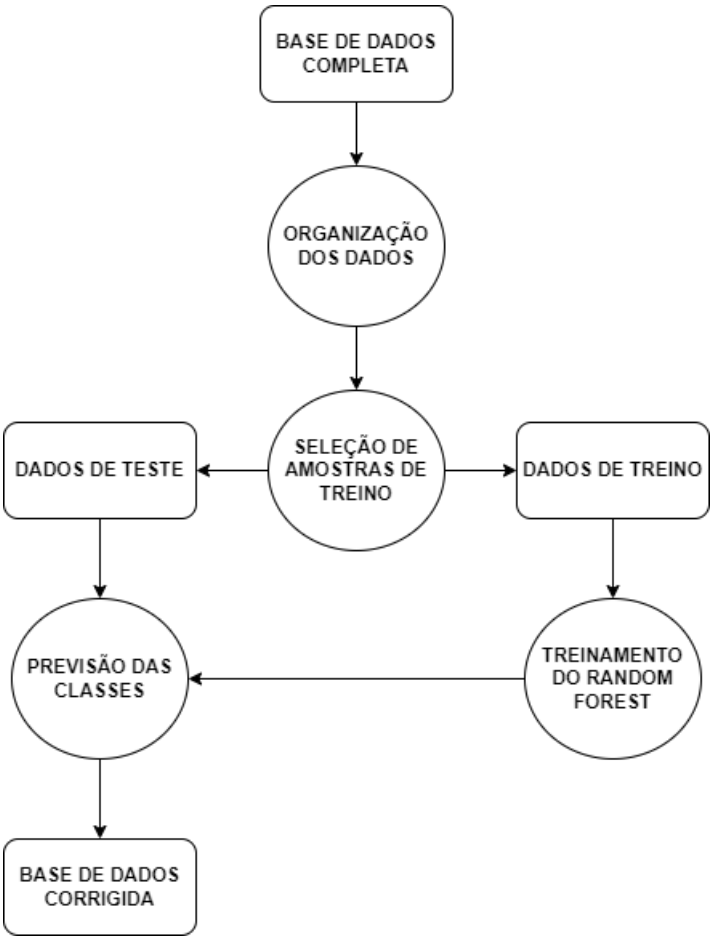

Figura 7- Fluxo de trabalho.

3.1 Base de dados

A base de dados contém a litologia descrita por testemunhos e a concentração de 16 elementos dados em porcentagem. As amostras de testemunho foram coletadas em um intervalo médio de 1 metro totalizando 12243 amostras que foram analisadas por fluorescência de raio-x nos laboratórios da SGSGeosol, para os elementos: silício (SiO2), alumínio (Al2O3), magnésio (MgO), sódio (Na2O), manganês (MnO), ferro (Fe2O3), titânio (TiO2), fósforo (P2O5), cálcio (CaO), potássio (K2O), cromo (Cr2O3), cobalto (Co), níquel (Ni), cobre (Cu), vanádio (V2O5) e bário (Ba). Os elementos Bário, Vanádio, Cobre e Sódio foram desconsiderados por falta de dados válidos. Além desses elementos contém informações como LOI (Loss on ignition) e Porcentagem total. O sumário estatístico desses dados encontra-se na Tabela 1. Foram analisados 737 furos (Figura 4) com distância variando de 100 a 200 metros entre si, sendo 528 de broca diamantada e 209 de circulação reversa.

Para a organização dos dados foram adotadas algumas medidas:

- Valores igual a zero foram substituídos por inválido.
- Valores abaixo do limite de detecção foram substituídos pela metade do limite.
- Valores negativos foram substituídos por abaixo do limite de detecção.
- Furos com valores inválidos foram desconsiderados.
- Intervalos com valores inválidos foram desconsiderados.
- Furos com litologias das encaixantes foram desconsiderados.
- Valores inválidos e abaixo do limite foram desconsiderados.

Assim, a base de dados foi organizada restando 12119 amostras. Na Figura 7 temos um esquema do fluxo de trabalho adotado.

	N Válidos	%Válidos	Média	Mediana	Mínimo	Máximo	Desv. Padrão
SIO2%	12243	100%	36,386	37,6	1,8	92,4	17,736
AL2O3%	12122	99,01%	6,104	4,5	0,05	29,9	5,381
MGO%	12243	100%	11,656	6,2	0,1	44	12,14
NA2O%	8504	69,46%	0,117	0,1	0,005	4,8	0,179
MNO%	12243	100%	0,512	0,37	0,03	12,6	0,456
FE2O3%	12243	100%	30,511	24,5	1	80,4	18,047
TIO2%	12243	100%	2,592	1,9	0,03	18,1	2,44
P2O5%	11722	95,74%	0,174	0,1	0,005	14,42	0,317
CAO%	10292	84,06%	1,787	0,14	0,005	25,8	3,681
K2O%	10276	83,93%	0,254	0,04	0,005	7,4	0,532
CR2O3%	12199	99,64%	1,534	0,95	0,005	12,9	1,517
LOI_%	12123	99,01%	7,956	8,31	0,005	38,28	3,001
CO%	11886	97,08%	0,049	0,03	0,003	0,717	0,05
CU%	8925	72,89%	0,014	0,01	0	0,22	0,014
NI%	12231	99,9%	0,479	0,306	0,005	3,733	0,455
V2O5%	9761	79,72%	0,186	0,06	0	17,1	0,969

Tabela 1. Estatística da base de dados.

3.2 Definição da estratigrafia do regolito

A estratigrafia do regolito foi definida de acordo com as descrições de campo utilizadas na época da exploração do depósito pela empresa Teck Cominco Ltd. Esses critérios foram os mesmos utilizados para o depósito de Ni laterítico de Santa Fé conforme reportado nos trabalhos de (Oliveira et al., 1992, Putzulo et al., 2020 e Machado, 2018). Por ser um depósito laterítico as litologias são de mesma idade geológica porém com ordem de deposição onde o primeiro horizonte a se formar é o R8. O depósito de Santa Fé localiza-se a aproximadamente 80 Km a N e com forte semelhança ao corpo do Rio dos Bois, pertencente à mesma Província Alcalina de Goiás. A seguinte sequência de horizontes do regolito foi definida para área do corpo do Rio dos Bois do topo a base:

- R1 – Solo pisolítico
- R2 – Ferricrete

- R3 – Zona de transição
- R4 – Silcrete
- R5 – Saprólito ocre
- R6 – Saprólito ferruginoso
- R7 – Saprólito verde
- R8 – *Saprock*
- U – Rocha ultramáfica fresca

Em razão do tratamento de dados alguns horizontes foram reunidos para simplificar a classificação pois representam classes geoquímicas similares. Os horizontes R1, R2 e R3 foram juntados em R123 (Figura 8); R5 e R6 foram juntados em R56.

Sendo assim, o resultado da junção das classificações dos horizontes intempéricos do regolito ficou:

- R123 – Zona superior do regolito
- R4 – Silcrete
- R56 – Zona intermediária oxidada
- R7 – Zona intermediária silicatada
- R8 – Zona do saprólito
- U – Rocha ultramáfica fresca

Figura 8- Amostras de testemunho mostrando os horizontes superiores. Fonte: Acervo do Professor C. Porto.

3.3 Machine Learning

Machine Learning surgiu em meados do século XX com o crescente uso de computadores, é de grande importância nas geociências atualmente principalmente por causa da crescente evolução tecnológica e da grande quantidade de dados disponíveis. Nos último anos trabalhos foram realizados utilizando o algoritmo Random Forest (Kortchmar, 2021 e Martins, 2021) ajudaram na compreensão do uso de machine learning e suas aplicações.

Para obter os resultados estatísticos necessários para atingir o objetivo do trabalho, o emprego do machine learning é essencial. Para tanto, foi empregado o uso de um algoritmo supervisionado denominado Random Forest Classifier, desenvolvido por (Breiman L., 2001).

Os algoritmos **supervisionados** têm o objetivo de classificar amostras baseadas em amostras previamente classificadas. As amostras já classificadas constituem uma base de dados de treino utilizada para treinar o modelo a identificar essas classes. Os algoritmos **não-supervisionados** classificam amostras que não têm uma classificação previa, criando assim, novos conjuntos de classe (Zhou, 2021).

O objetivo do emprego do algoritmo supervisionado é gerar um modelo que funcione em amostras não classificadas ou, como no caso desse trabalho, com classificações visuais de campo que estão em desacordo com sua composição geoquímica. Isto é feito utilizando amostras representativas de cada classe litológica para treinar o modelo. Este modelo serve para balizar as categorias geoquímicas dos horizontes intempéricos, assim, fazendo uma previsão das classificações de cada amostra baseado nas amostras de treino, cujas prévias classificações litológicas de campo são ideais já que foram confirmadas, por métodos estatísticos aplicados a seus dados geoquímicos, diminuindo assim o erro de generalização do modelo.

No estudo de machine learning existem certos conceitos para descrever fenômenos, como por exemplo, viés e variância. O **viés** mede a diferença entre a previsão do algoritmo e o real valor atribuído àquela amostra, expressando a habilidade de classificar corretamente. A **variância** mede a mudança de performance do algoritmo devido a mudanças nos dados de treino expressando o impacto dessa mudança no modelo. Outros conceitos como **overfitting** e **underfitting** compreendem o grau de generalização das previsões do dado algoritmo, onde o overfitting quer dizer que o modelo foca muito nas particularidades dos dados perdendo sua habilidade de generalizar e o underfitting quer dizer que o modelo foca muito nas generalizações e perde a habilidade de reconhecer pequenas diferenças.

3.3.1 Árvores de decisão

A árvore de decisão (Quinlan, 1993) é um algoritmo de machine learning que faz decisões baseado em uma estrutura de árvore. Os dados são divididos por meio da seleção de divisão para diferentes ramos, chamados de nós, com o objetivo de se obter apenas uma classe para cada nó no final do processo.

A sua estrutura (Figura 9) é seccionada em **nó pai** que é de onde a seleção de divisão começa e **nó filho** que é pra onde as amostras são direcionadas após a divisão de seleção. Um **nó pa**i gera **nós filhos** que vão virar **nós pais** e gerar mais **nós filhos** até atingir o objetivo de conter apenas uma classe dominante em um nó. A seleção de divisão controla para onde as amostras vão na árvore através da escolha de um atributo que as separa, sendo assim, é o mecanismo principal da árvore de decisão.

A impureza (Figura 10) pode ser calculada através da entropia e ganho de informação e isso controla a seleção de divisão de cada nó para cada um dos atributos do conjunto de dados. Cada **nó pai** corresponde a um teste de impureza para que um atributo de o mínimo de impureza nos **nós filhos** com o objetivo de não ter impureza para que a classe dominante seja única na seleção, assim, a árvore só acaba quando há uma classe dominante para cada nó.

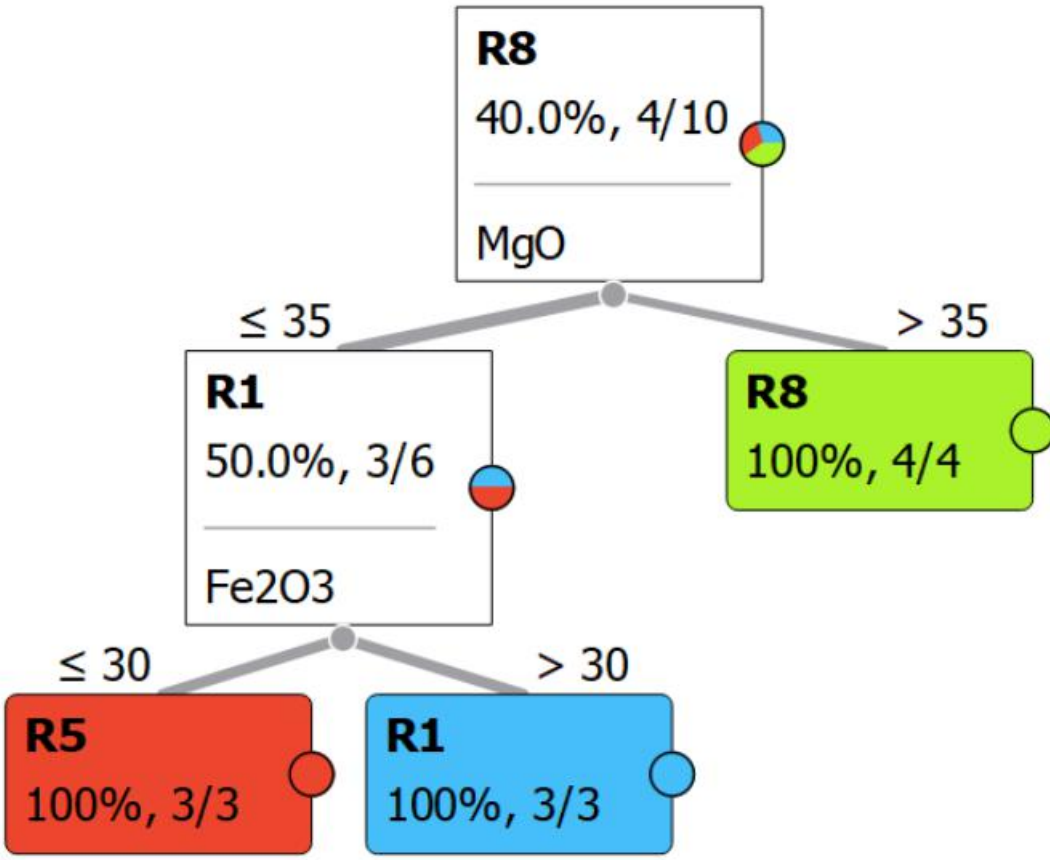

Figura 9- Árvore de decisão de dados de exemplo pelo software Orange Data Mining. A figura mostra a estrutura da Árvore de decisão e como se dá a seleção de divisão. O MgO é escolhido primeiro pois é o melhor atributo para separar as amostras entre classes de acordo com os calculos de impureza.

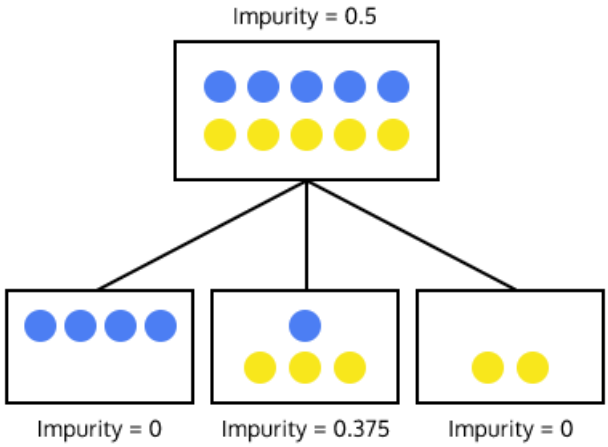

Figura 10: Esquema de distribuição de impureza. Fonte: Code Academy (https://www.codecademy.com/courses/machine-learning/lessons/ml-decision-trees/exercises/impurity)

Entropia

Entropia, diferente da termodinâmica, é uma função de impureza na teoria da informação. Ela é uma medida para saber o quão disperso são os dados em um conjunto observado (e.g. um atributo do conjunto de dados). A entropia diminui quando um conjunto de dados tende para uma única classe e aumenta quando tende para diversas classes. Para um conjunto de dados homogêneo a entropia é 0 e para um igualmente divido entre duas classes a entropia é 1.

O cálculo de entropia é dado da seguinte forma:

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k .$$

- Entropia do conjunto de dados D é igual a menos o somatório da multiplicação de Pk vezes o log base 2 de Pk para k igual 1 até |y|.
- Pk é a probabilidade da classe k (razão k/total de unidades em D) estar no conjunto de dados D, onde k = 1, 2, …, |y|.
- D é o valor alvo como por exemplo a classificação litológica dos horizontes intempéricos.

Quanto menor a entropia Ent (D) menor a impureza de D.

Ganho de informação

O ganho de informação é uma medida para saber o quanto de entropia se perde ao dividir um nó em uma árvore de decisão. Ele ajuda a escolher qual será a melhor coluna do conjunto de dados para que o nó pai nos dê nós filhos com o mínimo de impureza.

O cálculo de ganho de informação é dado da seguinte forma:

$$\mathrm{Gain}(D, a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v).$$

- Ganho de informação do conjunto de dados D em relação a um atributo *a* é igual a Entropia de D menos o somatório da razão Dv/D vezes a entropia de Dv, onde Dv é o conjunto de dados filho e v é o número de nós filhos que vai de 1 até v (D1, D2, …, Dv).
- Basicamente esta fórmula se traduz para: Ganho = Ent(pai) – SOMA Ent(filhos)

Então, esta fórmula nos dá uma medida de quanto de entropia os nós filhos terão para uma determinado atributo (i.e. MgO, Fe2O3, SiO2, etc) escolhida na seleção da divisão dos nós em respeito a variável alvo que seria a classificação litológica (R123, R56, R7, etc).

3.3.2 Random Forest

O algoritmo Random Forest (Breiman, 2001) consiste em um conjunto de múltiplas árvores de decisão que, antes da seleção de divisão, atributos são amostradas de forma aleatória similar ao método **bagging** para minimizar a variância. Então em geral, cada uma das árvores de decisão do Random Forest terá atributos (colunas) diferentes na hora da seleção de divisão do nó pai.

Com o uso deste algoritmo se perde a necessidade de normalização dos dados por lidar bem com diferentes unidades de medida e se objetiva minimizar a alta variância das árvores de decisão com o aprimoramento do método bagging. Segundo (Breiman, 2001*)*, o uso da Lei dos Grandes Números mostra que o Random Forest sempre converge para que overfitting não seja um problema. O fluxo de trabalho para o uso do Random Forest no software Orange Data Mining está representado na Figura 11.

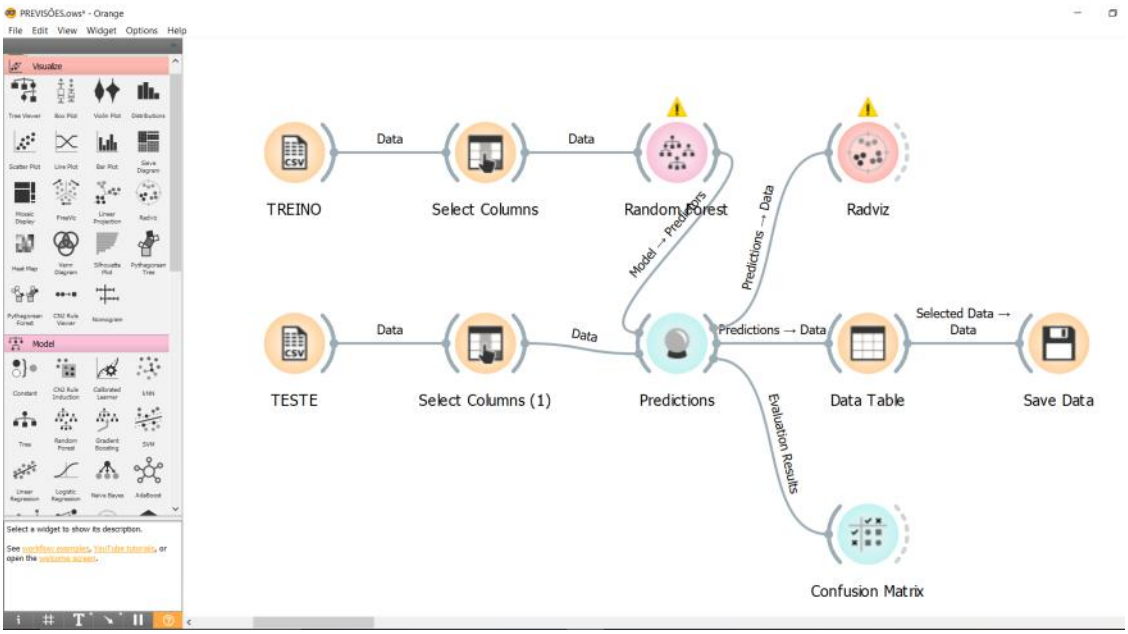

Figura 11: Fluxo de trabalho da previsão do Random Forest pelo software Orange Data Mining. Fonte : Acervo pessoal.

Boostrap

O boostrap (Efron e Tibshirani 1993) é uma técnica de amostragem onde a partir de um conjunto de dados D é criado um subset D' com o mesmo tamanho do original D porém com amostras repetidas e/ou faltando em relação ao original.

Bagging

O bagging (bootstrap aggregation) (Figura 12) é uma técnica de conjunto de aprendizagem (ensemble learning) onde se juntam diversos subsets, amostrados por meio de boostrap, diferentes para treinar em diversos algoritmos (learners) como, por exemplo, árvores de decisão e no final ter uma média dos resultados.

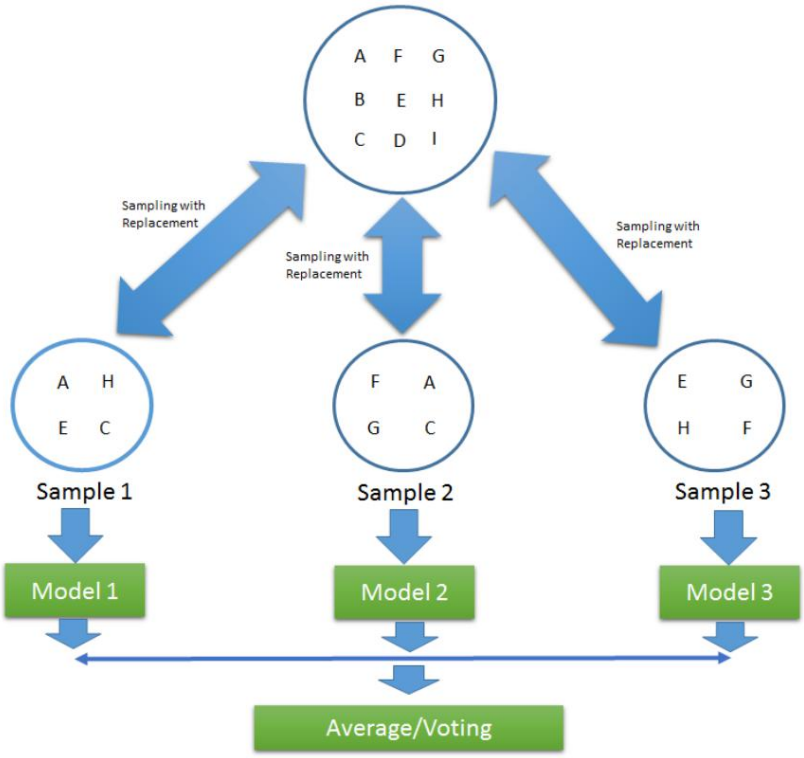

Figura 12: Esquema de como funciona o bagging. Começa com amostragem por boostrap e treina diferentes modelos para no final, obter uma média dos votos. Kumar, Rahul., 2019.

3.3.3 Treinamento

O treinamento dos dados foi feito com o intuito de balizar as classes litológicas através dos dados geoquímicos, sendo assim, foram utilizados os gráficos boxplots de cada elemento para selecionar as amostras entre o primeiro e terceiro quartil para todas as litologias de cada um dos quatro elementos principais. Estas amostras seriam as mais representativas de cada litologia e teriam os elementos com distribuição dentro do intervalo referido dos gráficos boxplots. Os óxidos escolhidos foram SiO2, MgO, Al2O3 e Fe2O3 que são os mais abundantes e assim mais impactam na definição das composições das classes litológicas.

Na Figura 13 está o gráfico boxplot do depósito Rio dos Bois mostrando a distribuição dos principais metais. A caixa azul representa a faixa de dados entre o 1º e o 3º quartil, a linha azul horizontal são os whisker, a linha vertical azul é a média e a amarela a mediana. Observou-se uma diferença na distribuição das amostras para estes elementos que facilitou a identificação e divisão das classes litológicas. Estas distribuições corroboram com as apresentadas nas referências bibliográficas como no trabalho de Elias 2002. Como exemplo, vemos o aumento dos teores de magnésio (Figura 13B) à medida que descemos no perfil para classes como R7 e R8 enquanto que os teores de ferro (C) diminuem. Estas distinções entre as classes nos permitiram selecionar as amostras que melhor representam sua distribuição geoquímica. Assim, foram selecionadas as amostras que estão dentro do intervalo entre o 1º e o 3º quartil dos quatro elementos para todas as classes.

Após a seleção das amostras de treino a base de dados (12119 amostras) foi divida entre Treino e Teste. As amostras do conjunto de dados de Treino (1356 amostras) foram utilizadas para treinar o modelo e as de Teste (10763 amostras) foram utilizadas para a previsão.

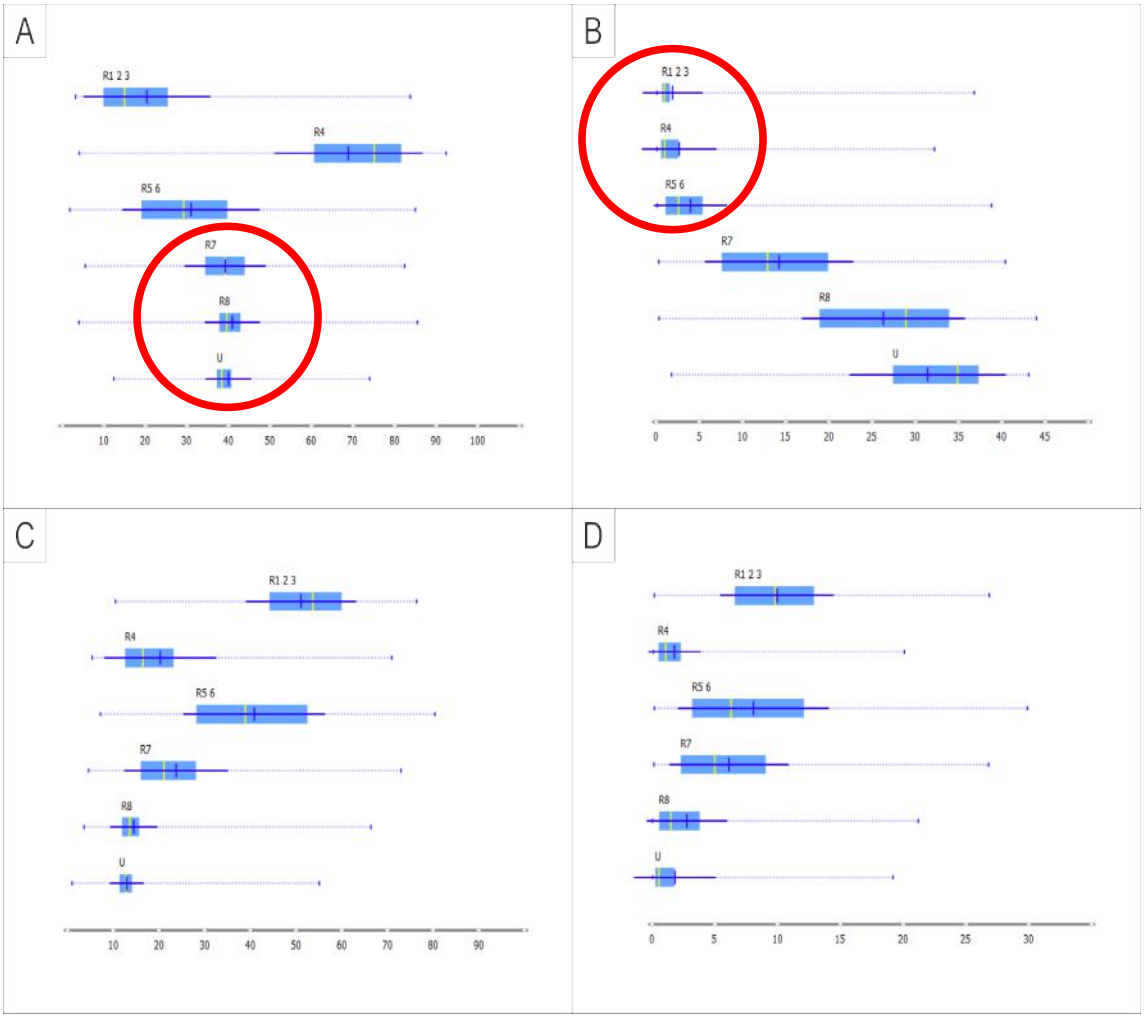

Figura 13- Gráfico boxplot de (A) sílica, (B) magnésio, (C) ferro e (D) alumínio para todas as classes litológicas das amostras do depósito do Rio dos Bois. Ressalta-se na imagem o comportamento das classes para certos óxidos. Eixo X com % em peso.

3.4 Visualização dos dados

Para grandes conjuntos de dados a visualização em duas ou três dimensões não consegue mostrar certas características dos conjuntos de classes plotados graficamente, sendo assim, sente-se a necessidade de se usar visualizadores multidimensionais para que o ser humano consiga alinhar bem o visual com o analítico e compreender melhor a relação entre os dados.

Utiliza-se então o RadViz proposto por Hoffman et al. (1997) com o objetivo de se visualizar em multidimensões dados numéricos com classificações distintas. Este método consiste em uma visualização radial que separa as dimensões em ancoras igualmente distribuídas no raio de um círculo, os valores das amostras são normalizados antes e a posição de cada ponto é relativa a resultante das forças de molas onde cada dimensão exerce uma força para a intensidade do seu peso no determinado ponto que por fim iguala essa resultante a 0 numa posição de equilíbrio.

Por exemplo, se todos os pontos têm os mesmos valores eles ficaram no centro do círculo, se uma amostra tem alta intensidade em uma dimensão ela ficará mais perto desta ancora.

Utilizou-se também gráficos como boxplots, gráficos de frequência e matrizes de confusão para melhor entendimento da disposição geoquímica e litológica dos dados.

4 RESULTADOS E DISCUSSÃO

Foram gerados dois modelos diferentes para comparação de qual seria a melhor previsão para as classes da base de dados. Os dois modelos foram treinados pelas mesmas amostras evidenciadas no **tópico 3.3.3** e a previsão foi realizada no conjunto de dados de Teste, porém, com diferentes variáveis consideradas pelo Random Forest. O **modelo 1** considerou os 4 metais principais formadores das litologias (SiO2, MgO, Fe2O3 e Al2O3). O **modelo 2** considerou 13 atributos sendo eles 12 elementos (SiO2, MgO, Fe2O3, Al2O3, CaO, K2O, MnO, TiO2, P2O5, Cr2O3, Co e Ni) e Perda ao Fogo.

Para os parâmetros do Random Forest foram utilizadas 100 árvores de decisão sem limite de crescimento onde todos os atributos do conjunto de dados são considerados em cada seleção de divisão. Também foi necessário manter o balanceamento de classes para preservar a proporção da distribuição das litologias.

Ressalta-se que a previsão do Random Forest pode gerar inúmeras previsões, e foi observado que a cada previsão havia uma variação da distribuição de classes pois o algoritmo funciona a base de votos que a cada previsão variam. Sendo assim, podem haver pequenas diferenças entre modelos que podem ser interpretados como amostras que são de difícil classificação para o algortimo por estarem em zonas transitórias de classes geoquímicas.

4.1 Gráficos RadViz

Na Figura 14 estão plotadas todas as amostras do depósito Rio dos Bois. Observa-se que a distribuição geoquímica das amostras do depósito se sobrepõe em grande parte deixando pouco claro o que cada classe representa geoquimicamente. A classe R56 (verde) é a mais abundante e está presente no setor de outras três classes R123, R4 e R7 apenas não sobrepondo as classes R8 e U.

Para o conjunto de dados de Treino (Figura 15) observa-se que as classes litológicas se separam em setores geoquímicos distintos onde o R123 (azul) situa-se em um extremo perto do ferro e o U (lilás) no outro extremo perto do magnésio tendo ainda uma pequena sobreposição por parte das litologias R123 com R56 e R8 com U. Essas sobreposições se devem ao fato de que a transição entre R123 e R56 assim como R8 e U é muito gradativa, e também porque podem ocorrer mais de uma classe litológica dentro do intervalo médio de 1 metro do testemunho descrito e analisado e, por fim, por que as fases minerais são remobilizadas por estruturas, como fraturas e falhas, e condicionamentos intempéricos que dificultam as distinções de classe.

Após a previsão do Random Forest, com as amostras de teste, para o modelo 1 e 2 como visto na Figura 16 e Figura 17, observa-se que ainda há sobreposições entre as classes, porém as mesmas encontram-se melhor definidas em suas zonas e a sobreposição da classe R56 (verde) se reduz consideravelmente, podendo assim, classificar melhor as amostras que caiam em campos sobrepostos. Os dois modelos são semelhantes apenas com pequenas mudanças.

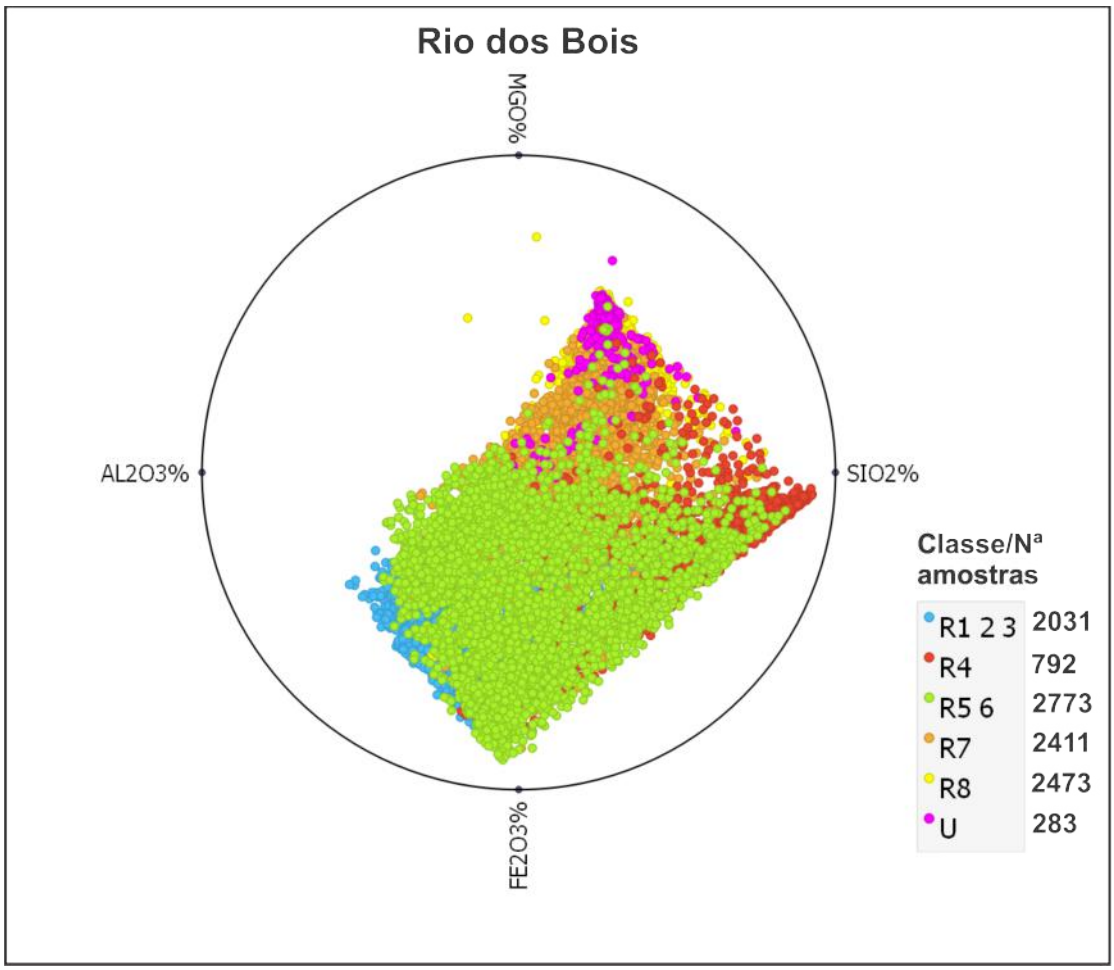

Figura 14: Gráfico RadViz de todas amostras do depósito Rio dos Bois com número de amostras para cada classe.

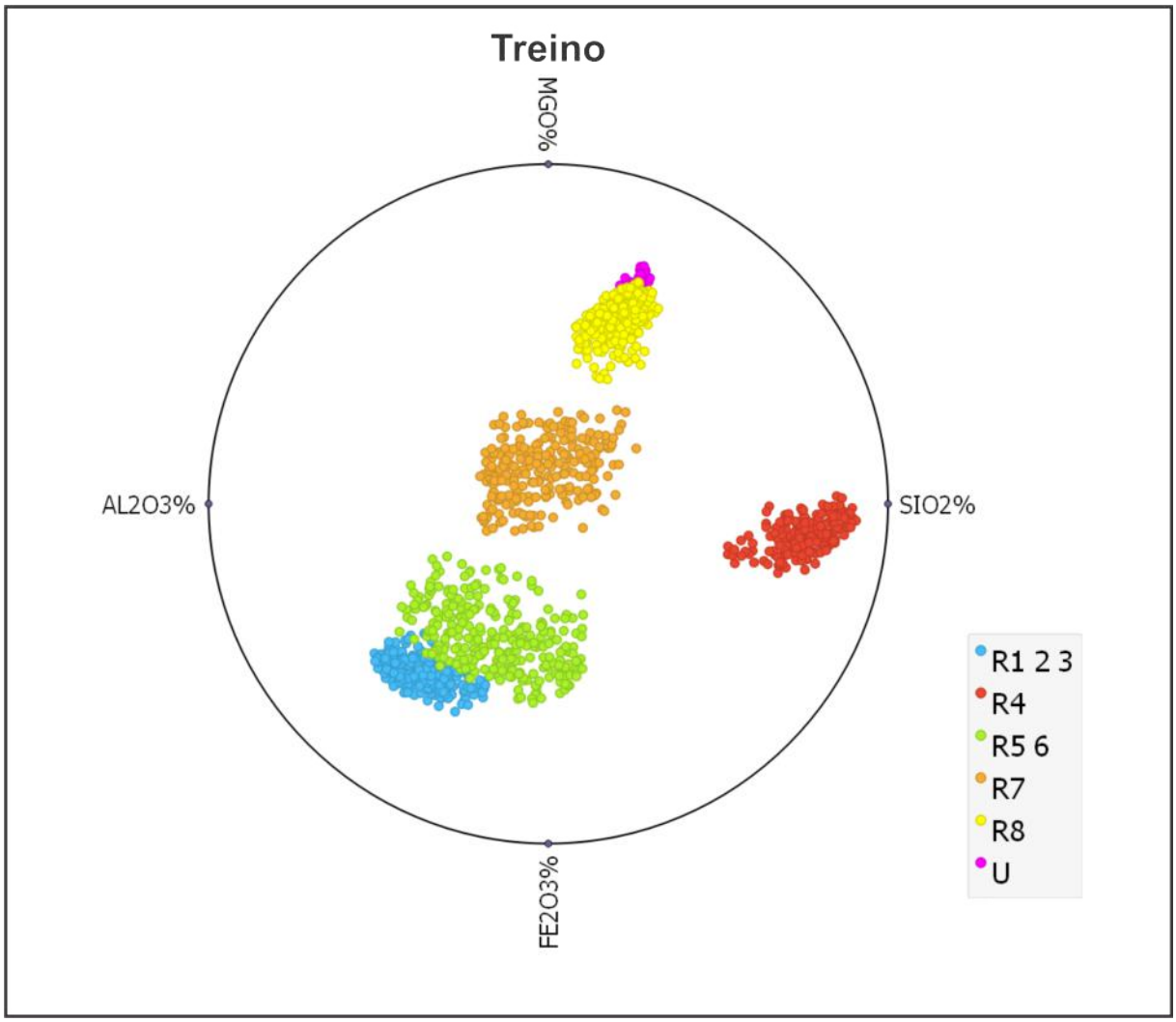

Figura 15: Gráfico RadViz das amostras do conjunto de dados Treino com suas respectivas classes.

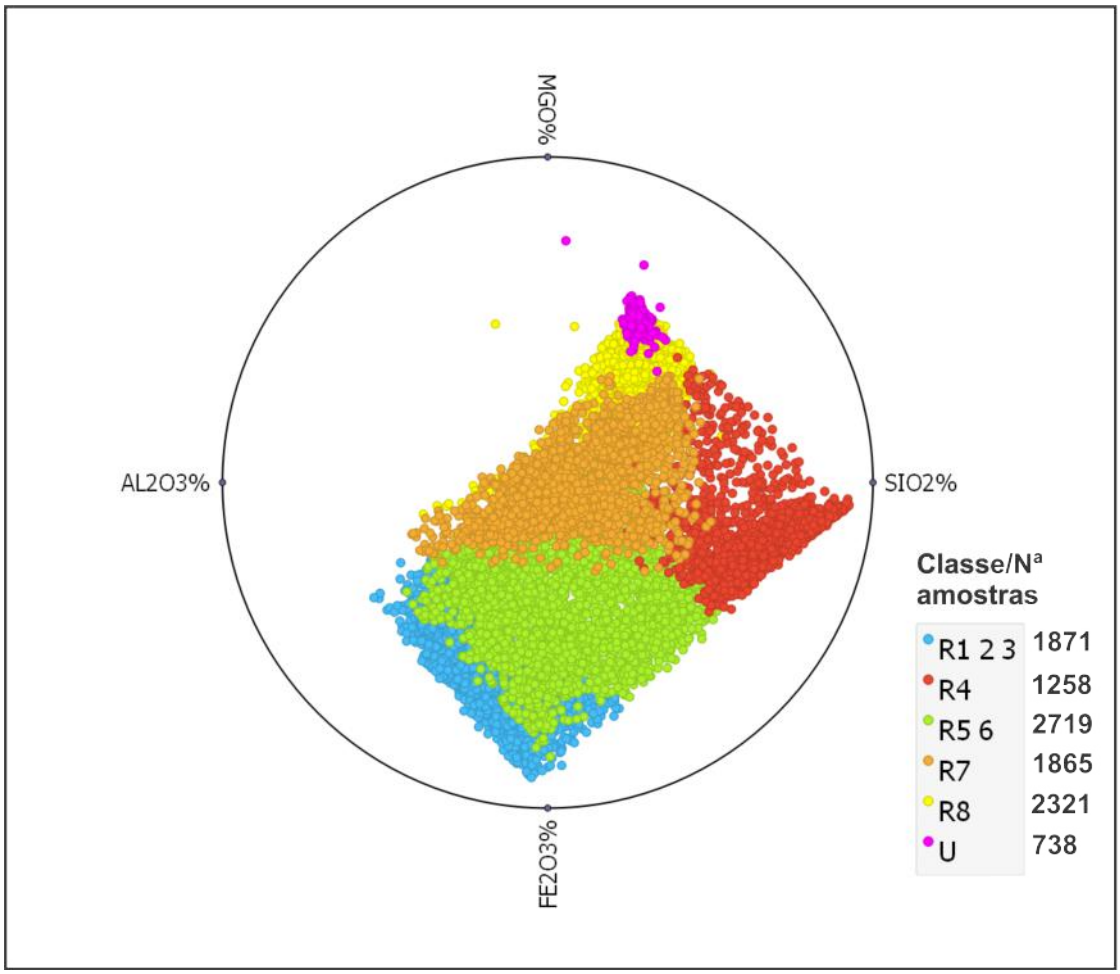

Figura 16: RadViz das amostras do conjunto de dados Teste após a previsão para o modelo 1 com número de amostras para cada classe.

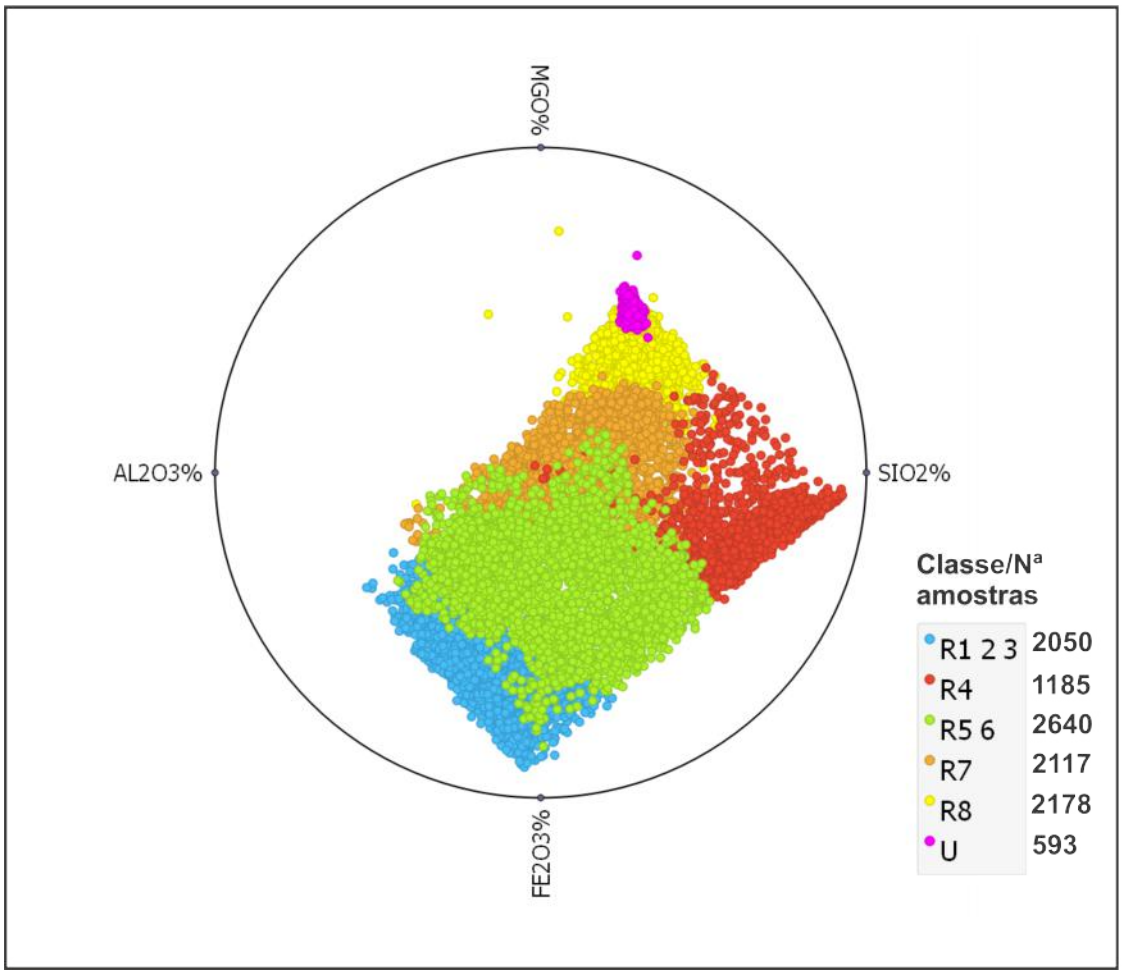

Figura 17: RadViz das amostras do conjunto de dados Teste após a previsão para o modelo 2 com número de amostras para cada classe.

4.2 Gráficos de frequência

Para os gráficos de frequência, vemos na Figura 18 o eixo y representando a quantidade de amostras e o eixo x representando a classe litológica do conjunto de dados Teste antes da previsão do Random Forest. Nota-se as proporções de ocorrência das classes litológicas. As classes R7 e R8 apresentam números parecidos de frequência e as classes U e R4 poucas ocorrências.

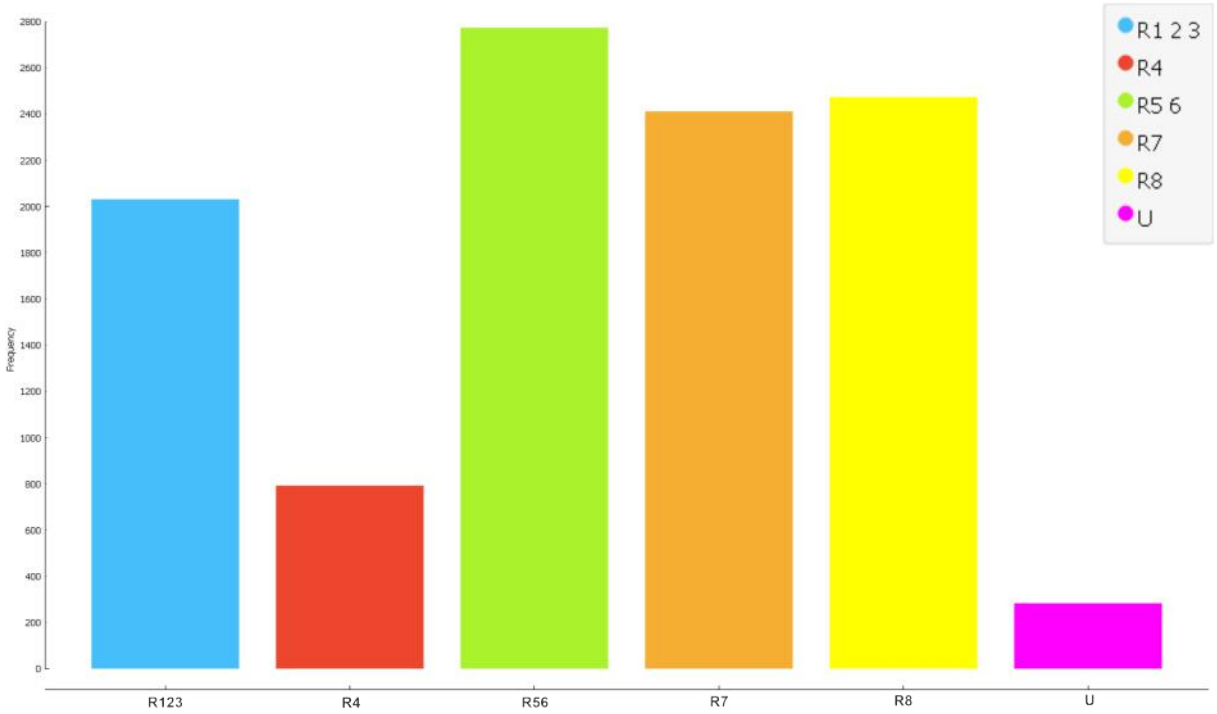

Figura 18: Gráfico de frequência das amostras do conjunto de dados Teste antes da previsão do Random Forest. O eixo y representa a quantidade de amostras e o eixo x representa a classe litológica.

Na Figura 19 vemos a distribuição de classes após a previsão do Random Forest para o modelo 1, observa-se um aumento da classe U que pode ser por dois fatores. Primeiro, pela dificuldade do algoritmo entender a diferença química entre R8 e U. Segundo, pela difícil separação das classes pois o R8 contém diversos blocos de dunito (U) imersos na matriz do regolito. Pode ser pelos dois fatores ou por simplesmente ter mais U e a previsão estar correta. Nota-se também que o R7 diminui e o R4 aumenta em dimensões de cerca de 400 amostras cada um. Na Figura 20 temos a distribuição de classes após a previsão pro modelo 2. Pode-se ver que há apenas uma leve diferença do R7 pro R8 em relação ao modelo 1, não evidenciando ainda nenhuma diferença importante entre os modelos.

Nas Figuras 21 e 22 (modelo 1 e 2 respectivamente) vemos que o eixo x e a quantidade de amostra das barras são correspondentes às classes litológicas antes da previsão e a cor corresponde as classes depois da previsão, podendo assim, ver para quais classes as amostras foram reclassificadas. Observou-se que uma quantidade perecida de amostras de R123 foram reclassificadas para R56 e de R56 para R123. Estas duas classes têm uma zona de interferência que dificulta a distinção exata delas, podendo ter as duas em um intervalo de 1 metro. Quase um quarto do R8 original foi reclassificado para U. R4 e R56 estavam presentes em quase todas as classes litológicas.

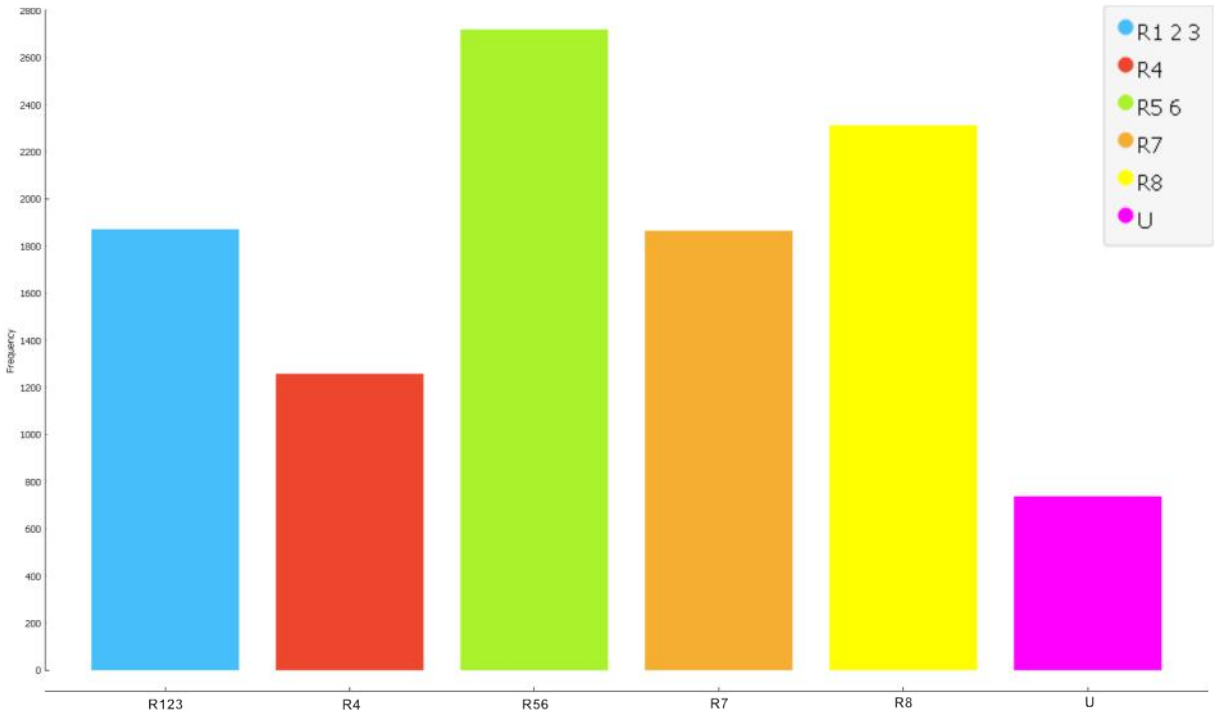

Figura 19: Gráfico de frequência das amostras do conjunto de dados Teste após previsão para o modelo 1. O eixo y representa a quantidade de amostras e o eixo x representa a classe litológica.

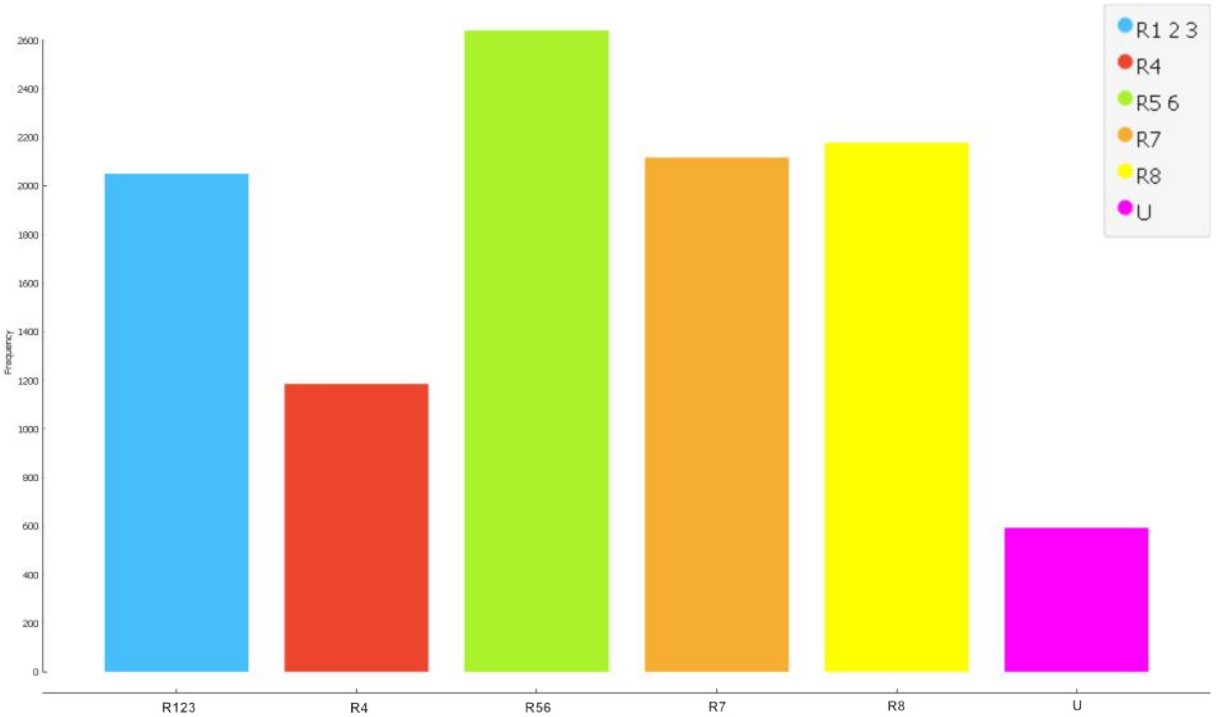

Figura 20: Gráfico de frequência das amostras do conjunto de dados Teste após previsão para o modelo 2. O eixo y representa a quantidade de amostras e o eixo x representa a classe litológica.

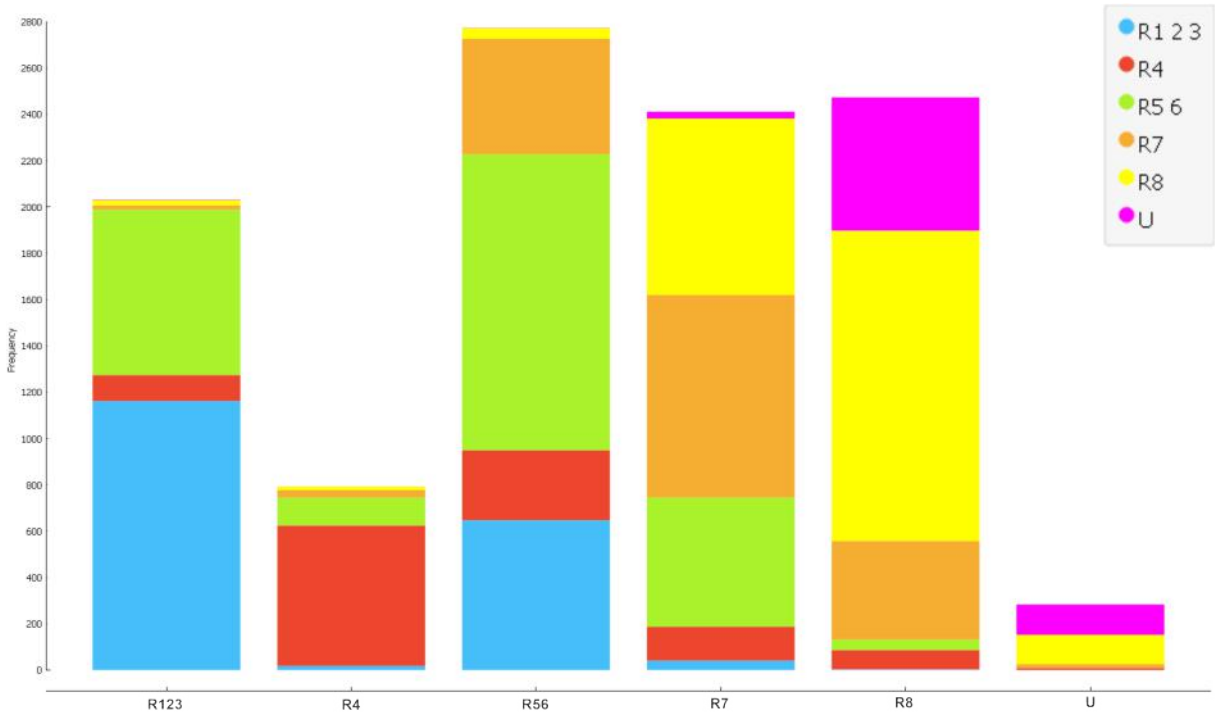

Figura 21: Gráfico de frequência das amostras do conjunto de dados Teste após previsão para o modelo 1. O eixo y representa a quantidade de amostras, o eixo x representa a classe litológica antes da previsão e as cores representam as classes depois da previsão.

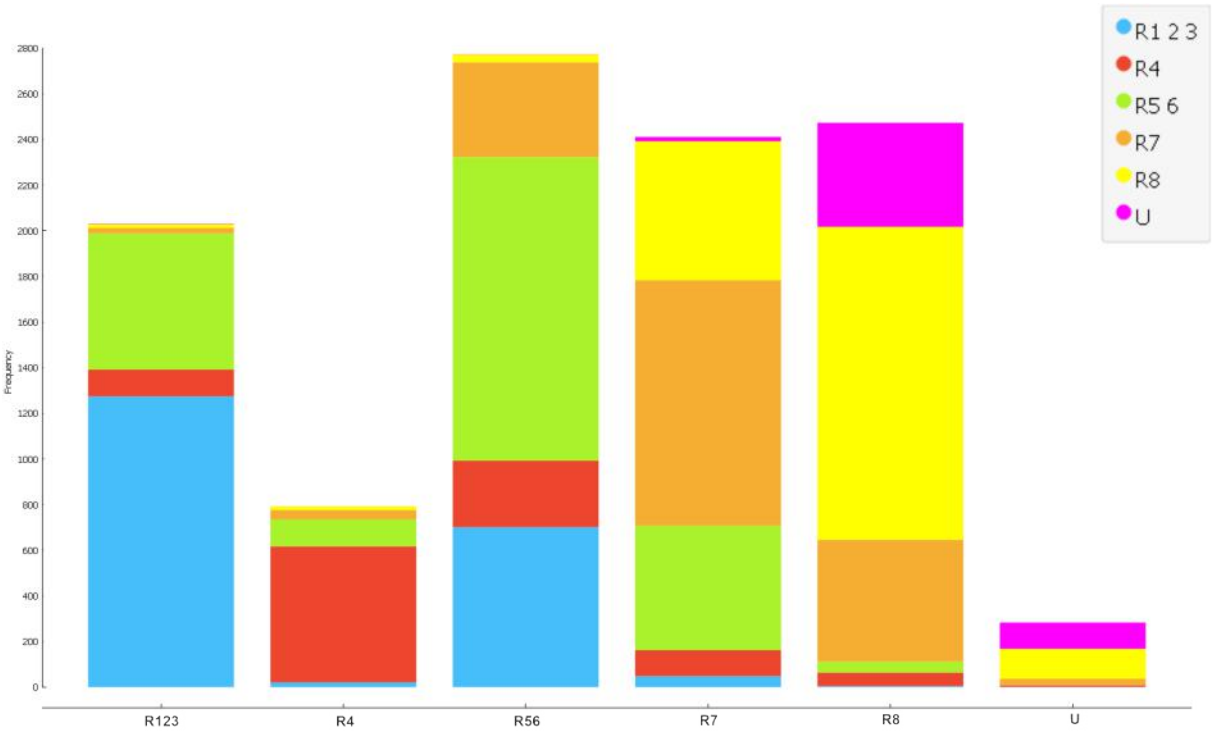

Figura 22: Gráfico de frequência das amostras do conjunto de dados Teste após previsão para o modelo 2. O eixo y representa a quantidade de amostras, o eixo x representa a classe litológica antes da previsão e as cores representam as classes depois da previsão.

4.3 Matrizes de confusão

As matrizes de confusão nos informam da quantidade de amostras que estavam presentes nas classes antes e depois da previsão. No eixo x estão as amostras antes da previsão, suas classes e a soma do total de amostras. No eixo y estão as amostras previstas pelos modelos. Elas complementam as informações descritas na seção anterior (4.2) onde podemos ver a quantidade exata de amostras reclassificadas e suas classes. Os modelos 1 e 2 estão presentes nas Figuras 23 e 24. A cor azul representa as amostras que continuaram na sua classe e a vermelha são as que foram reclassificadas. A intensidade da cor corresponde a quantidade de amostras.

647 amostras do R123 original viraram R56 e vice-versa para os dois modelos. Como visto na seção anterior R123 e R56 apresentam uma zona de interferência com muitas amostras e isso ocorre com outras classes também. Para as classes R56 com R7 e R7 com R8 também há bastante amostras reclassificadas. Isso se deve aos fatores já comentados que podem ser amostras de transição entre litologias. Os dois modelos se comportam muito parecidos com algumas flutuações pequenas na dimensão de cerca de 100 amostras (1% do total de amostras).

Como mencionado anteriormente, as previsões do Random Forest variam um pouco, então, esta diferença na proporção das classes pode ser interpretada apenas como uma dessas variações, onde temos algumas amostras que são de difícil classificação pelo algoritmo por representarem zonas geoquímicas transitórias entre classes. Por isso os dois modelos vão apresentar pequenas variações.

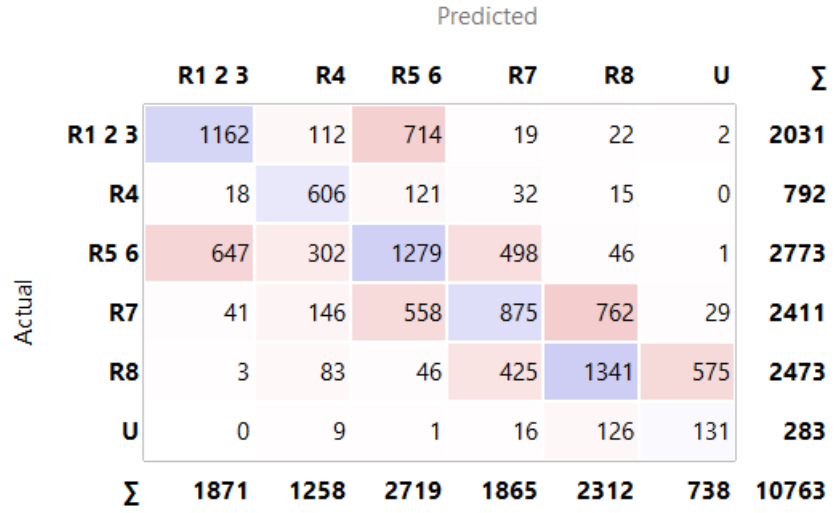

Figura 23- Matriz de confusão para o modelo 1.

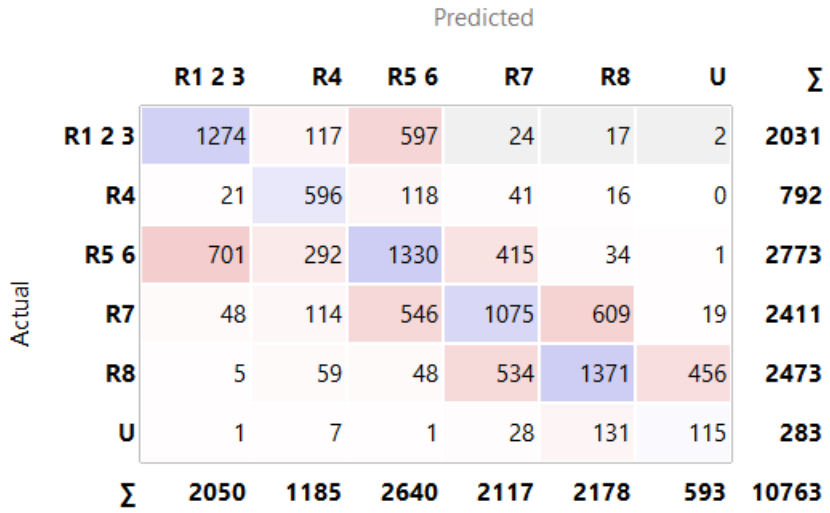

Figura 24- Matriz de confusão para o modelo 2.

4.4 Gráficos boxplot

Foram gerados 4 gráficos de boxplot compreendendo os 4 óxidos de metais principais que compõe as amostras: Si, Mg, Fe e Al. Os boxplots da Figura 25 mostra a distribuição dos dados do Teste antes da previsão. Nas Figuras 26 e 27 vemos os boxplots para amostras previstas dos modelos 1 e 2. Nota-se que esses últimos mostram uma distribuição mais uniforme das amostras para as classes em relação a Figura 25. É evidente a diminuição da dispersão dos dados para sílica (A), ferro (C) e para magnésio (B) em especial para a classe R123, mostrando o desempenho do algoritmo de reclassificar as amostras para seus devidos grupos geoquímicos. Para os dois modelos houveram mudanças expressivas na dispersão. No magnésio para a classe U foi observado uma boa diminuição na dispersão dos dados. E no alumínio as classes R56 e R7 aumentaram e diminuiram respectivamente suas médias. Não houveram variações expressivas entre os modelos.

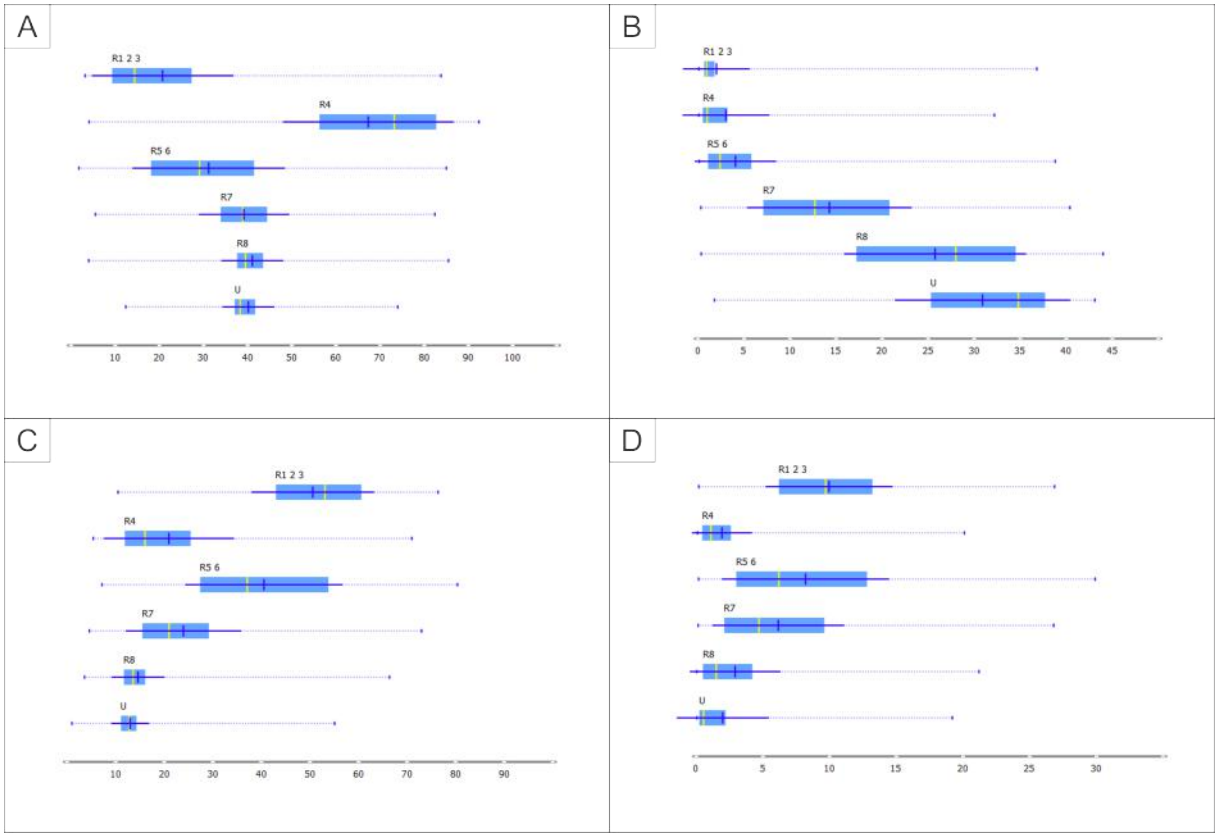

Figura 25: Gráfico boxplot de (A) sílica, (B) magnésio, (C) ferro e (D) alumínio para todas as classes litológicas das amostras do conjunto de dados Teste antes da previsão.

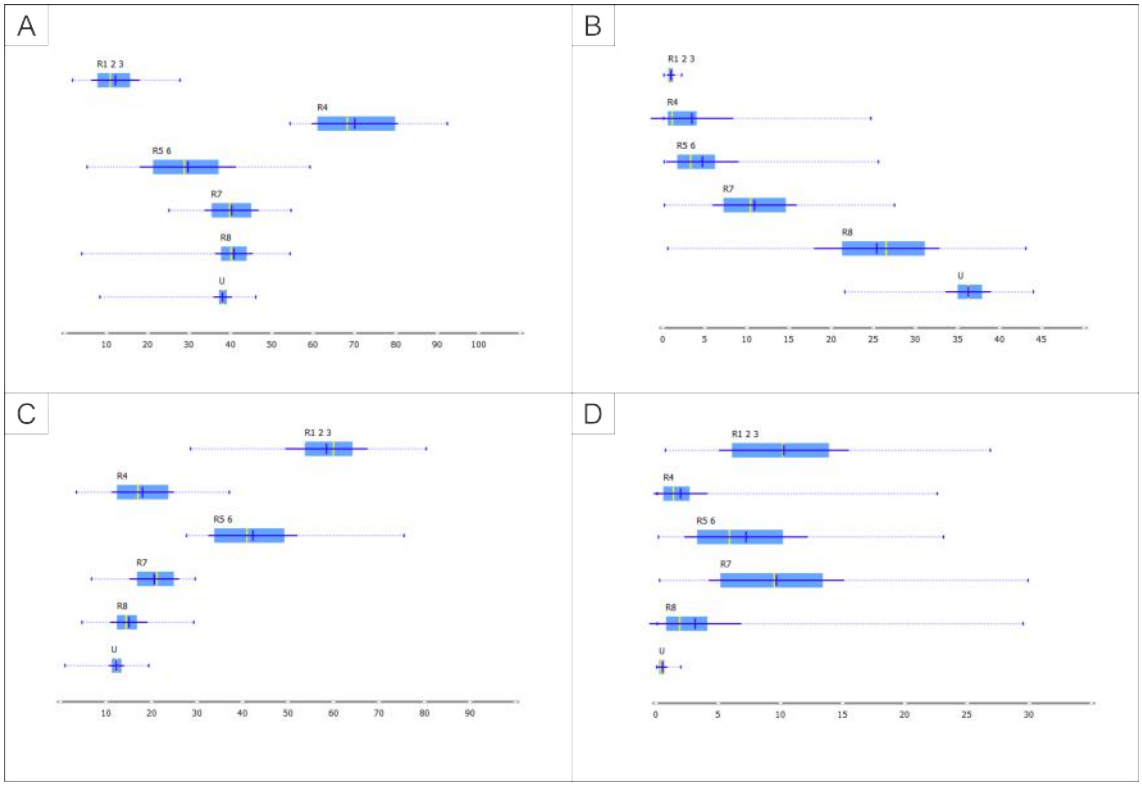

Figura 26- Gráfico boxplot de (A) sílica, (B) magnésio, (C) ferro e (D) alumínio para todas as classes litológicas das amostras previstas para o modelo 1.

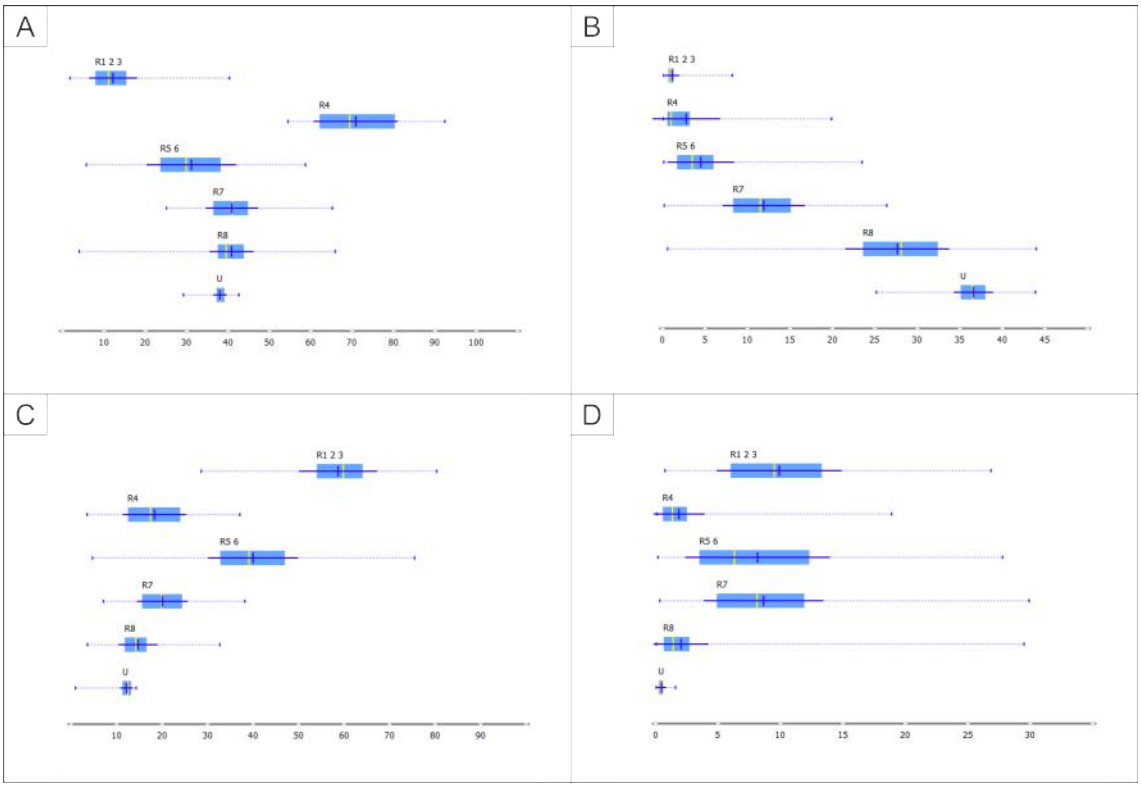

Figura 27- Gráfico boxplot de (A) sílica, (B) magnésio, (C) ferro e (D) alumínio para todas as classes litológicas das amostras previstas para o modelo 2.

4.5 Distribuição do minério

Após a previsão dos dois modelos foi feita uma análise da distribuição das classes por meio dos gráficos de frequência para amostras que estão acima do teor de corte de 1.0% de Ni e 0.1% de Co. Assim, podemos ver quais são as mudanças de volume de minério para todas as classes previstas pelo Random Forest.

Na Figura 28 temos a distribuição de minério com teor de corte 1.0% Ni de todas as classes para as amostras do conjunto de dados Teste. As Figuras 29 e 30 são referentes as amostras previstas do modelo 1 e 2. Tem-se que as mudanças mais relevantes após a previsão são a diminuição de amostras consideradas minério pelo teor de corte de 1.0% Ni da classe R7 e um aumento de amostras da classe R123 para os dois modelos, porém, as mudanças mais expressivas foram para o modelo 2. Estas mudanças nos levam a entender que houve uma mudança no volume do tipo de minério diminuindo o silicático hidratado e aumentando o oxidado. O modelo 2 reclassifica ainda mais amostras do R56 para o R123.

De acordo com (Golithly, 2010), a classe R123 não é uma boa portadora das fases minerais que hospedam níquel. O gráfico boxplot de Ni para amostras do depósito Rio dos Bois (Figura 31) também evidencia isso mostrando uma média de 0,38% Ni. Então, a maior quantidade de amostras consideradas minério que foram reclassificadas para R123 podem ser por outros motivos, como a incapacidade do Random Forest de separar certas amostras de R123 do R56, onde que para o R56 a média do teor foi a mais alta com 0.75% Ni. Foram analisadas as médias de profundidade no furo das amostras consideradas minério com teor de corte 1.0% Ni para o conjunto de dados de Teste antes da previsão (Figura 32) e depois da previsão para os dois modelos (Figura 33 e 34) por meio dos gráficos boxplots. Vê-se que a média da profundidade das amostras da classe R123 aumenta consideravelmente após a

previsão para os dois modelos nos levando a entender que possivelmente esse aumento na quantidade de amostras acima do teor de corte que foram reclassificadas para R123 foram confundidas com amostras da classe R56 que tem média de profundidade um pouco maior. Apesar da inclusão do níquel no modelo 2 junto dos outros atributos não houveram diferenças expressivas em relação ao modelo 1. Outra possibilidade é de que a classe R123 realmente tem mais amostras com teor de Ni acima do teor de corte utilizado e o algoritmo evidenciou elas.

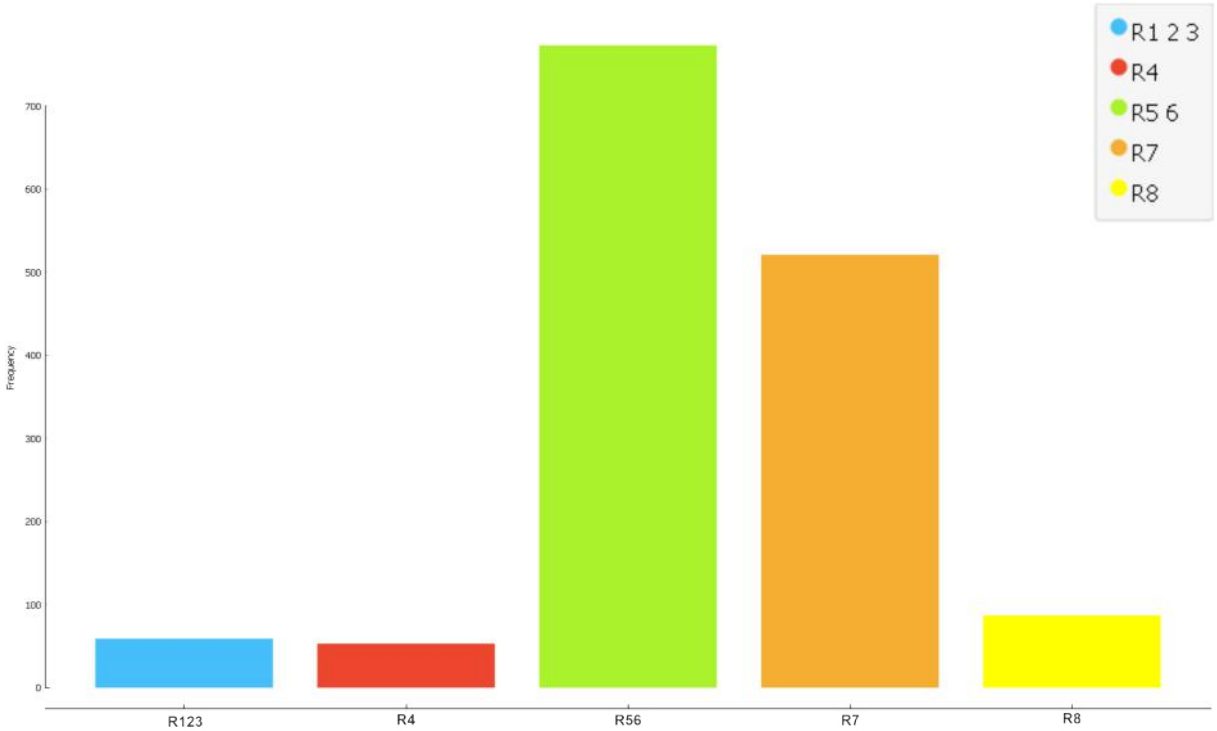

Figura 28: Gráfico de frequência para número de amostras acima do teor de corte 1.0% Ni em cada classe do conjunto de dados Teste.

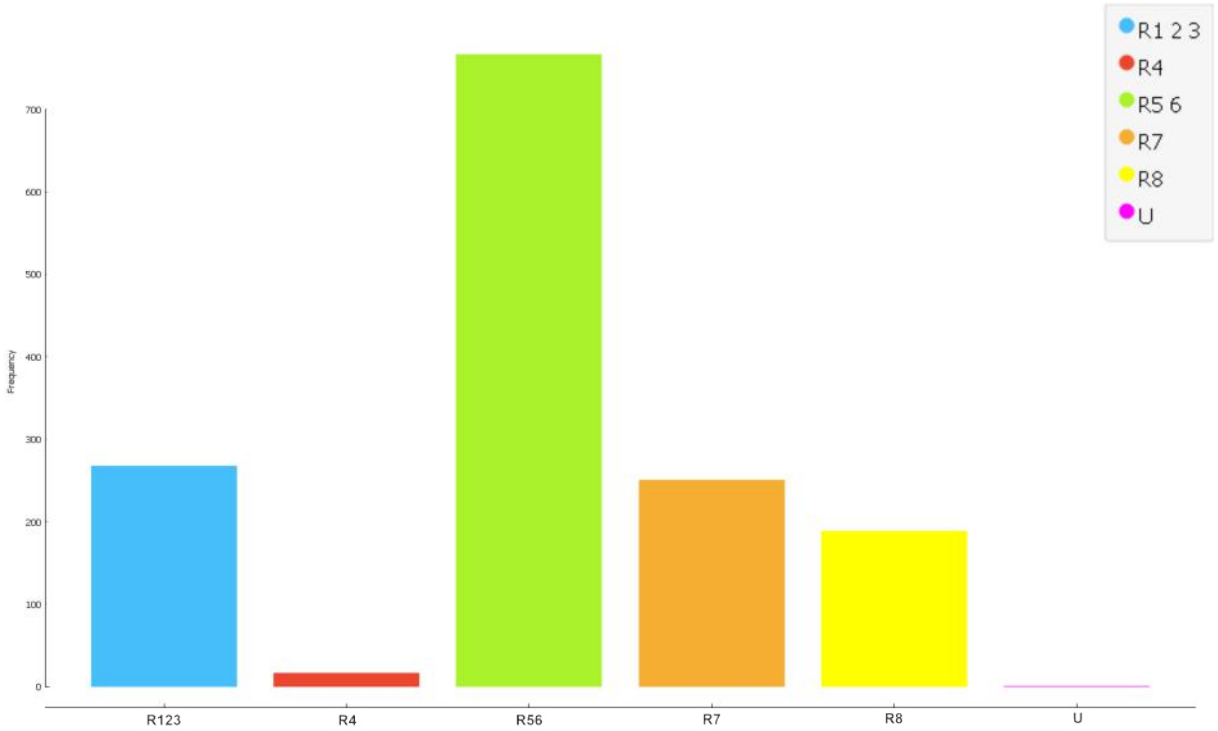

Figura 29: Gráfico de frequência para número de amostras acima do teor de corte 1.0% Ni em cada classe do conjunto de dados previstos para o modelo 1.

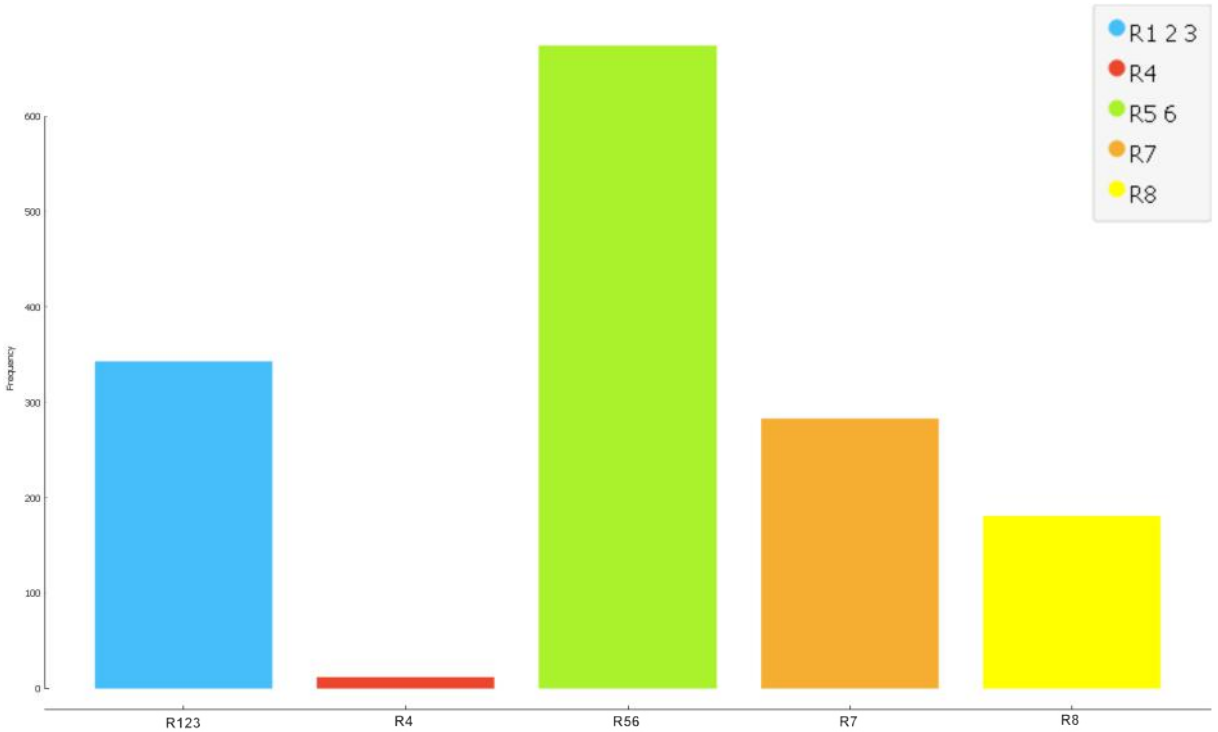

Figura 30: Gráfico de frequência para número de amostras acima do teor de corte 1.0% Ni em cada classe do conjunto de dados previstos para o modelo 2.

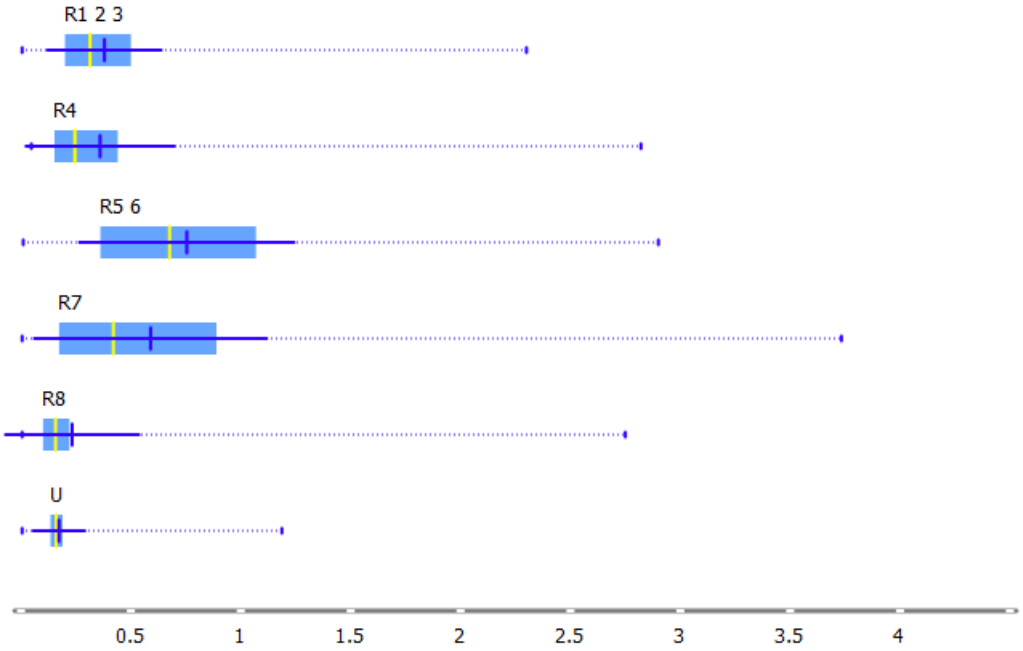

Figura 31- Gráfico boxplot de Ni para amostras do depósito Rio dos Bois.

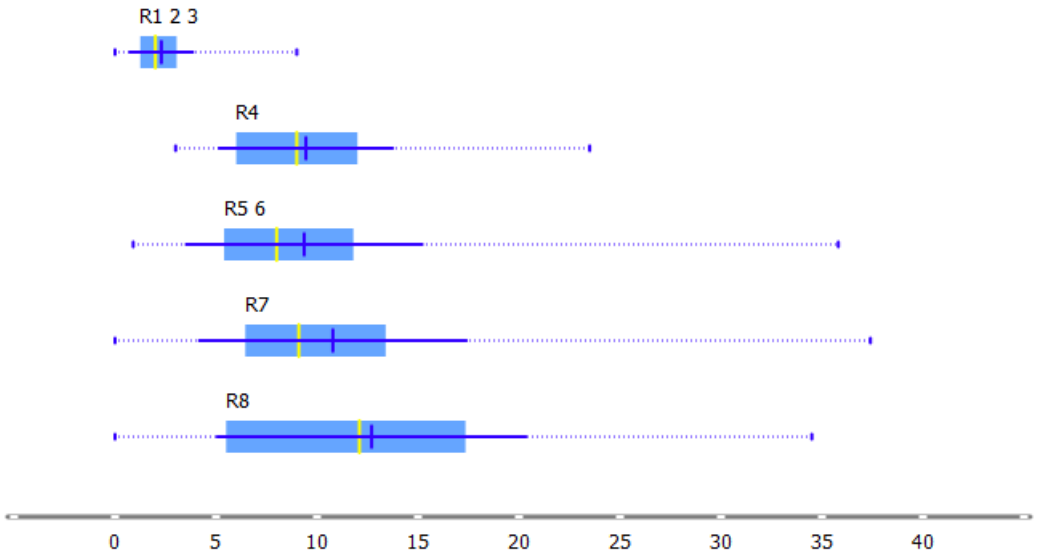

Figura 32- Gráfico boxplot de profundidade em metros para amostras acima do teor de corte de 1.0% Ni do conjunto de dados Teste.

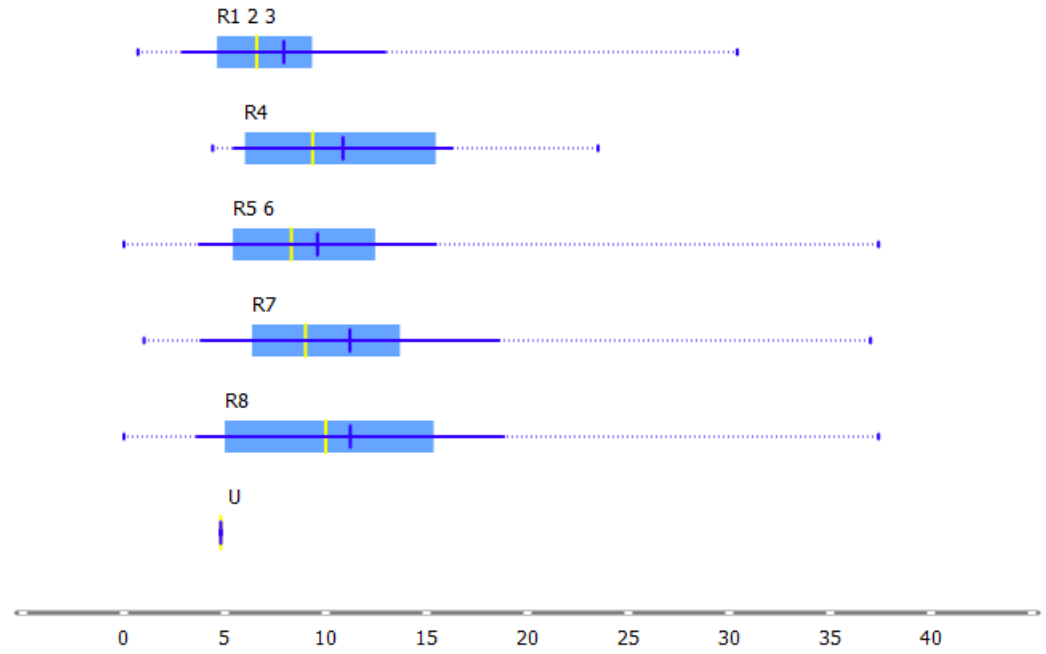

Figura 33: Gráfico boxplot de profundidade em metros para amostras acima do teor de corte de 1.0% Ni previstas para o modelo 1.

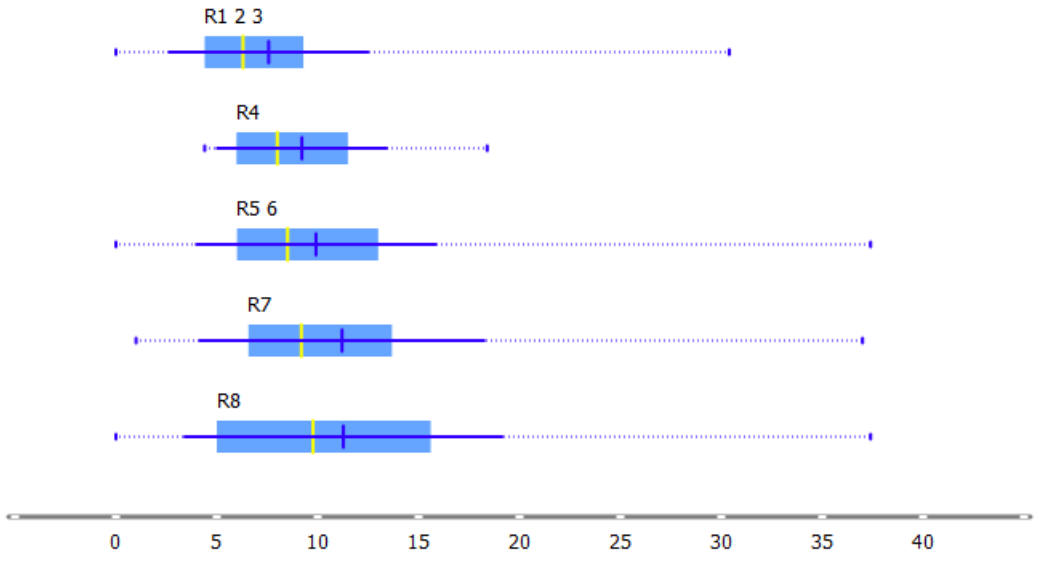

Figura 34: Gráfico boxplot de profundidade em metros para amostras acima do teor de corte de 1.0% Ni previstas para o modelo 2.

Na Figura 35 temos a distribuição de minério com teor de corte 0.1% Co de todas as classes para as amostras do conjunto de dados Teste. Nas Figuras 36 e 37 vemos as amostras previstas do modelo 1 e 2. De primeira nota-se que as amostras da classe R56 diminuem consideravelmente e as da classe R123 aumentam após a previsão do Random Forest. Parece coerente que as amostras reclassificadas para R123 preencham mais o espaço de minério de colbalto que provavelmente eram amostras da classe R56 que pela quantidade de cobalto foram reclassificadas. Neste caso o Random Forest parece entender bem a divisão entre as amostras da classe R123 e R56 para teor de cobalto, pois a classe litológica R123 hospeda os maiores teores do depósito como visto na Figura 38 onde vemos o boxplot de Co para as classes do depósito Rio dos Bois antes da previsão.

Temos que a média (linha vertical azul) do teor de Co (Figura 38) antes da previsão para a classe R123 é de cerca de 0.089% Co e para classe R56 é 0.072% Co, após a previsão os dois modelos (Figura 39 e 40) apresentaram 0.1% Co para R123 e 0.068 para R56. Este aumento na média do teor de cobalto pode ser explicado pela boa reclassificação de algumas amostras que anteriormente eram em sua maioria R56 e foram reclassificadas para R123.

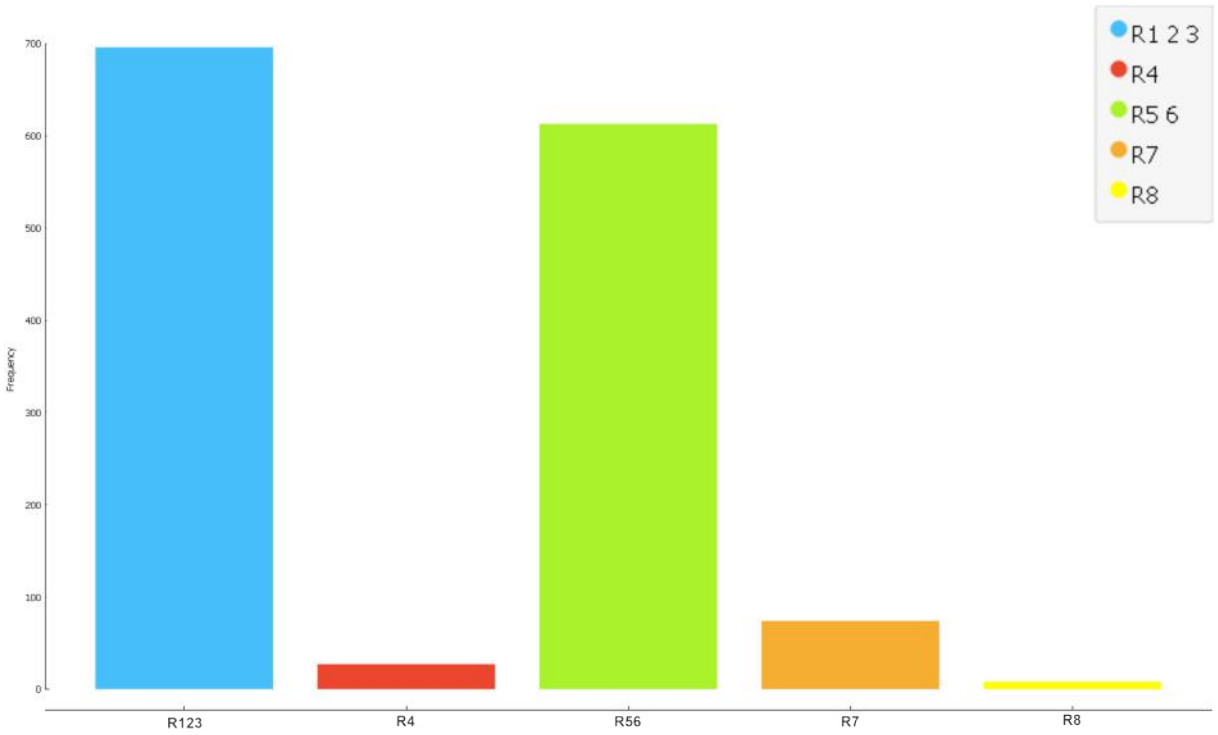

Figura 35: Gráfico de frequência para número de amostras acima do teor de corte 0.1% Co em cada classe do conjunto de dados Teste antes da previsão.

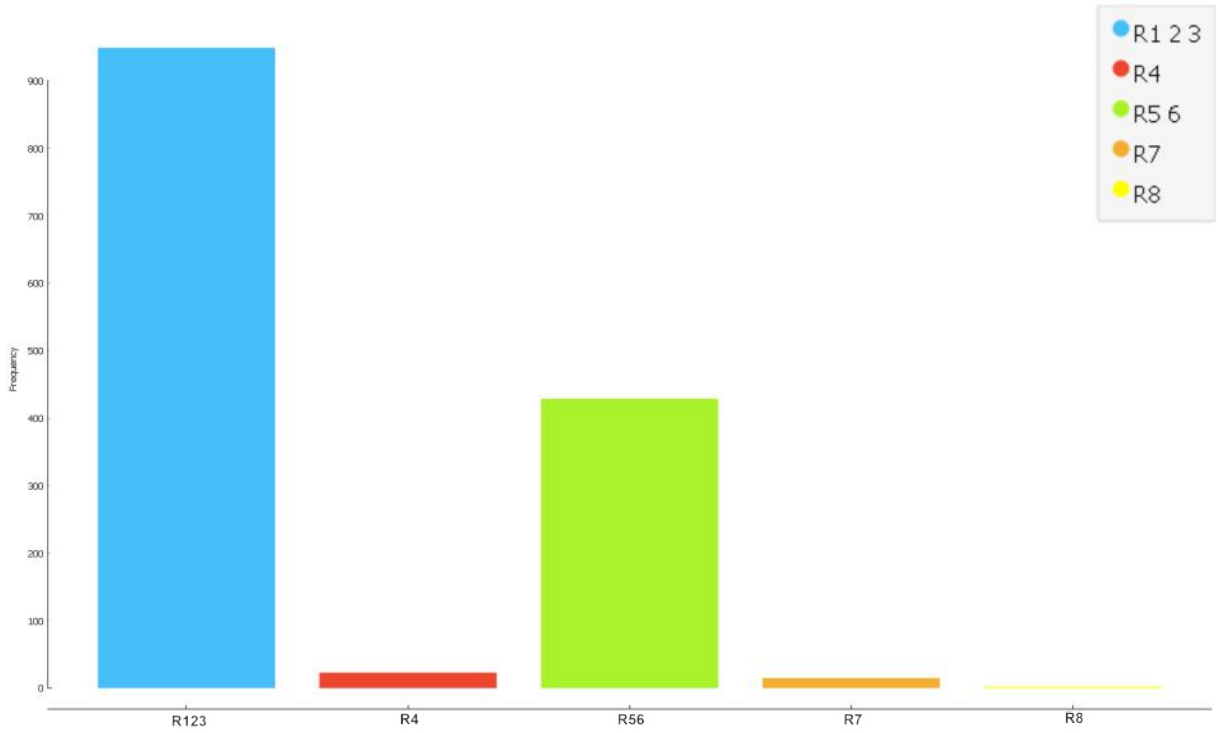

Figura 36: Gráfico de frequência para número de amostras acima do teor de corte 0.1% Co em cada classe do conjunto de dados previstos para o modelo 1.

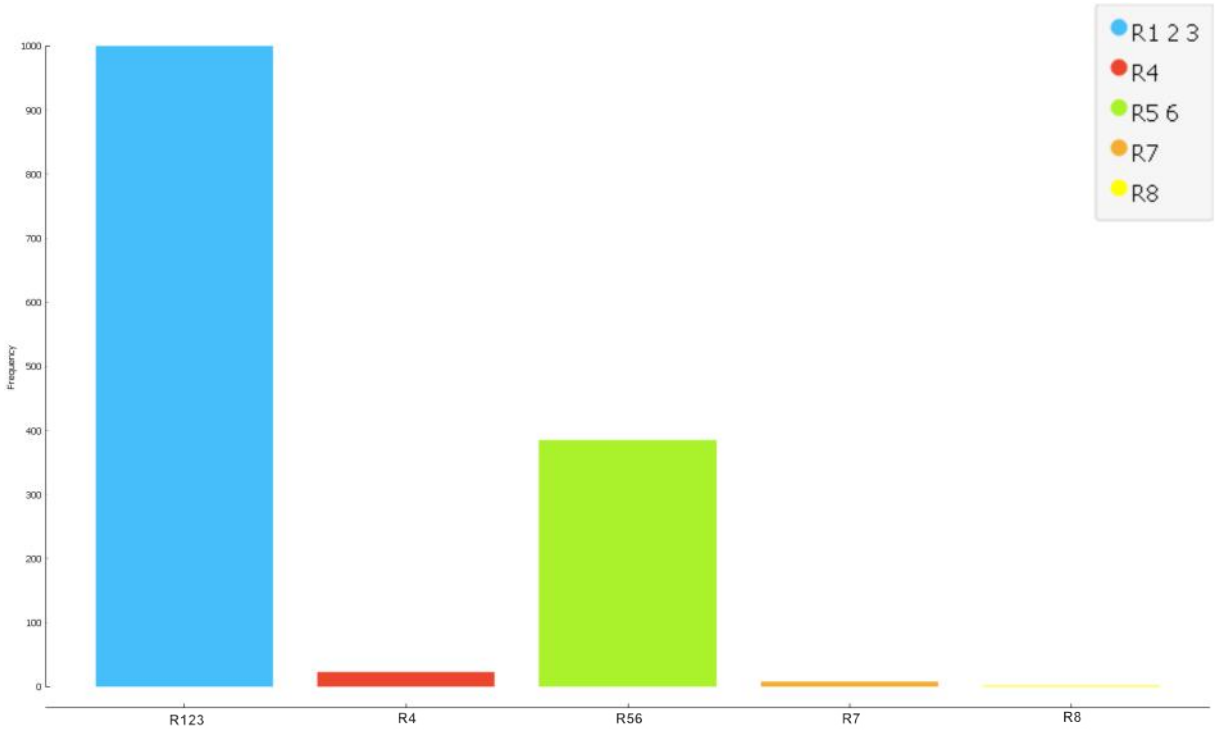

Figura 37: Gráfico de frequência para número de amostras acima do teor de corte 0.1% Co em cada classe do conjunto de dados previstos para o modelo 2.

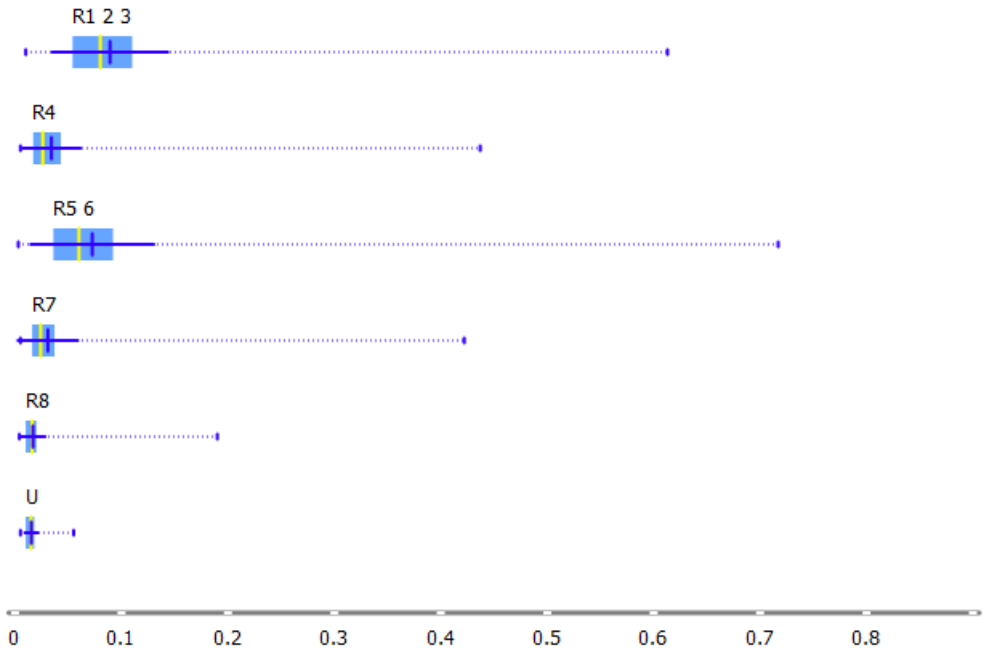

Figura 38: Gráfico boxplot de Co para amostras do depósito Rio dos Bois.

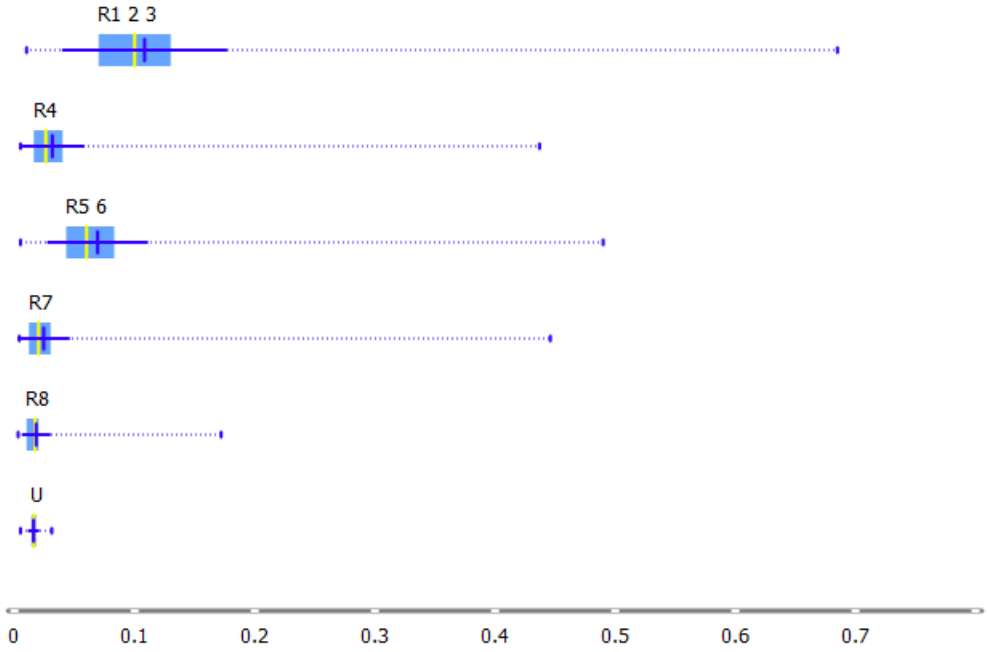

Figura 39: Gráfico boxplot de Co para amostras previstas para o modelo 1.

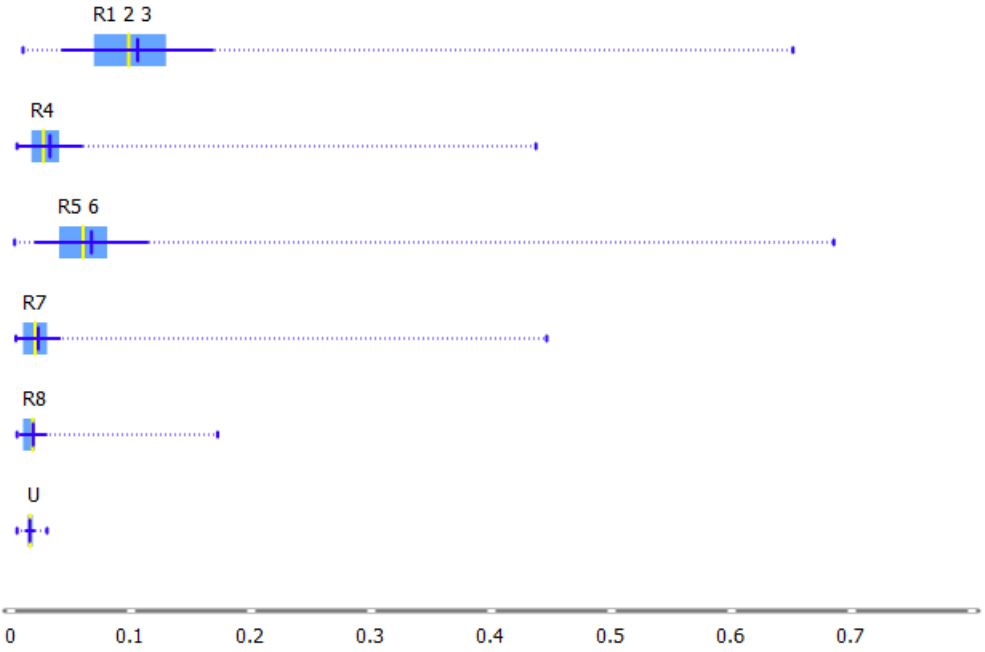

Figura 40: Gráfico boxplot de Co para amostras previstas para o modelo 2.

4.6 Mapas de horizontes mineralizados

Na Figuras 41 e 42 estão os mapas de elevação mostrando as amostras dos modelos 1 e 2 que estão acima do teor de corte de 1.0% Ni. A cor lilás representa amostras de minério oxidado (R123 e R56) que foram reclassificadas para minério silicatado (R7 e R8). A cor azul representa o contrário, minério silicatado que foi reclassificado para oxidado.

Observa-se primeiro, uma predominância de ocorrências de amostras de cor azul na porção norte do depósito onde a elevação é moderada. Nas porções sul e oeste ocorrem mais amostras de cor lilás onde a elevação do terreno é mais alta. Talvez a previsão do Random Forest nos mostre um padrão na formação desses depósitos em relação a sua posição no terreno. A parte sul está localizada na borda de um morro acarretando um espessamento do perfil e uma maior lixiviação propiciando melhor desenvolvimento das classes R7 e R8. Já as amostras da classe R123 são menos desenvolvidas devido à maior declividade do terreno. Na parte norte o terreno plano concede melhor desenvolvimento dos horizontes superiores.

A mudança na espacialidade dos tipos de minério é muito importante para diversas etapas da exploração mineral.

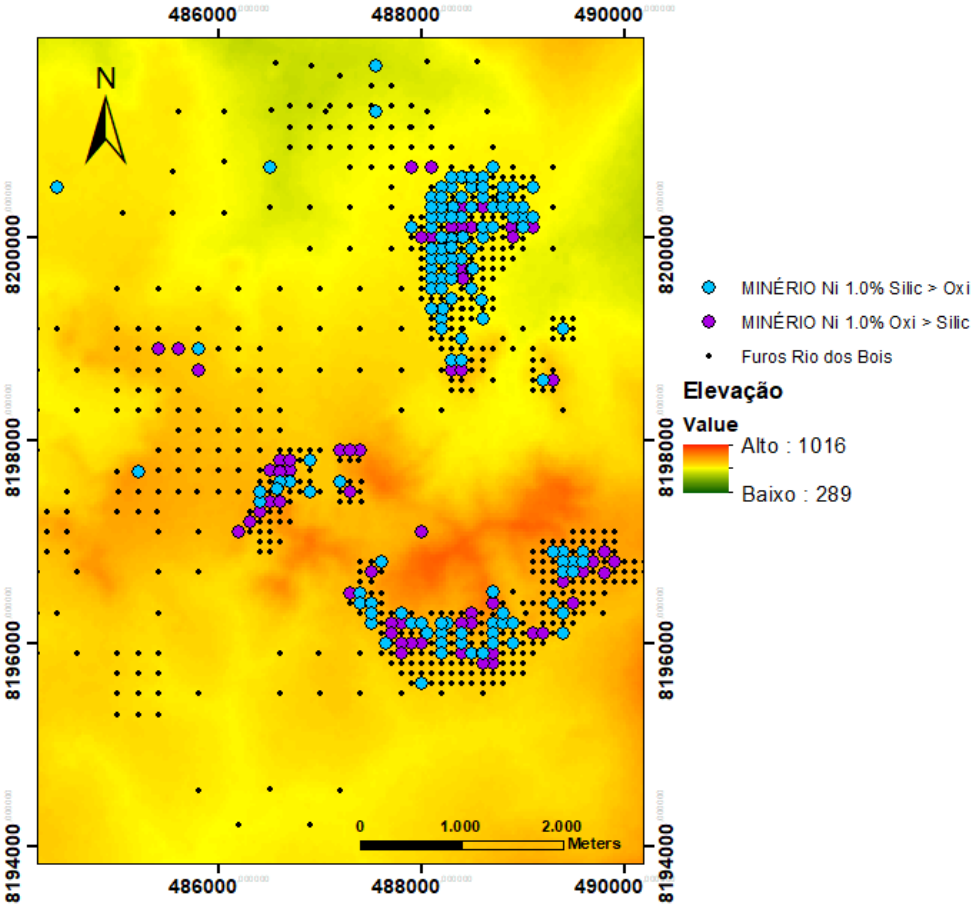

Figura 41- Mapa de elevação com amostras previstas para o modelo 1. A cor lilás representa amostras de minério oxidado (R123 e R56) que foram reclassificadas para minério silicatado (R7 e R8). A cor azul representa o contrário, minério silicatado que foi reclassificado para oxidado.

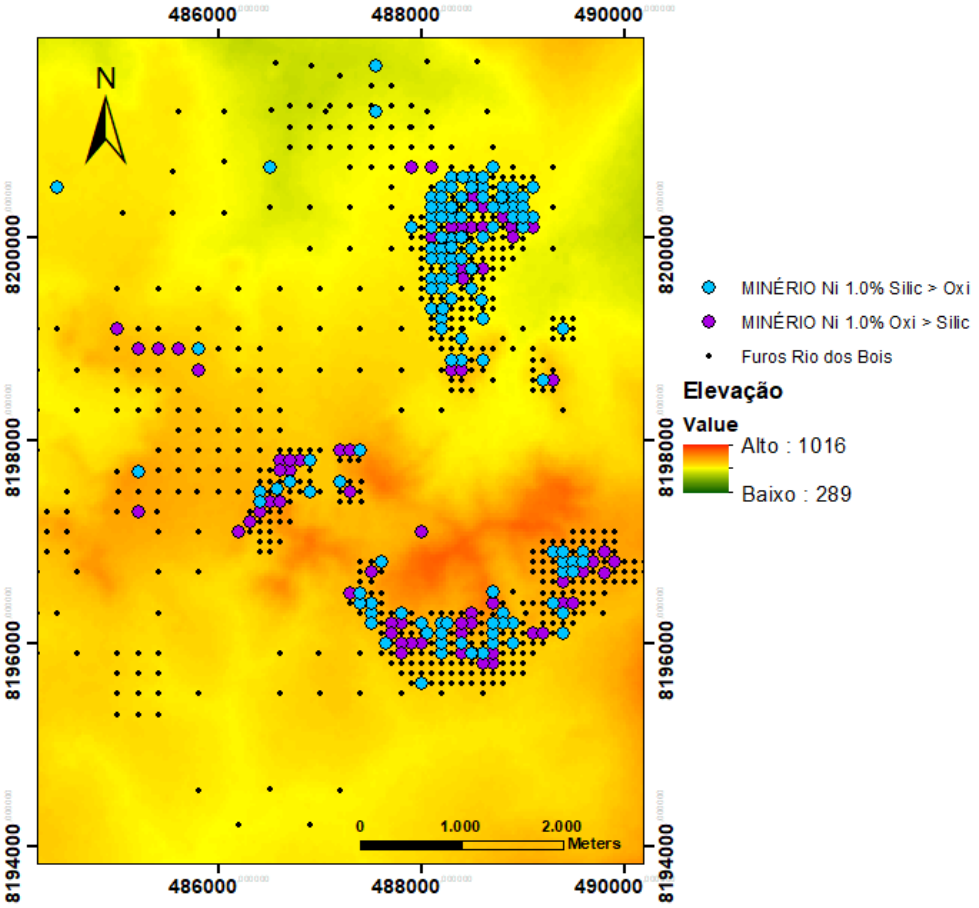

Figura 42- Mapa de elevação com amostras previstas para o modelo 2. A cor lilás representa amostras de minério oxidado (R123 e R56) que foram reclassificadas para minério silicatado (R7 e R8). A cor azul representa o contrário, minério silicatado que foi reclassificado para oxidado.

5 CONCLUSÃO

O emprego de machine learning em banco de dados se mostra eficiente para diferentes tipos de análises, principalmente quando se integra às criteriosas observações geológicas geradas no campo. Com os devidos objetivos, mostra-se um bom empenho na manipulação de dados estatísticos para correlacionar diversos tipos de dados para um melhor entendimento dos casos.

Os modelos previstos nos ajudam a entender melhor a disposição geoquímica das amostras com suas classes, podendo evidenciar informações acerca das fases minerais presentes nas litologias do depósito. No caso de minérios, ajudou a entender melhor a relação de proporção entre os minérios oxidados e silicatados. Também ajudam para compreender as relações geoquímicas entre as classes e os diferentes fatores geológicos que podem empregar papéis fundamentais no desenvolvimento do perfil lateríticos e de suas mineralizações.

De acordo com os resultados de testes realizados com a inclusão de um maior número de variáveis como no modelo 2, o Random Forest produziu resultados bem semelhantes ao modelo 1 que usa apenas 4 variáveis. Houve pequenas variações na classificação de algumas amostras comparando os dois modelos mas entende-se que o algoritmo enfrenta difíceis escolhas na análise de amostras que não se encaixam exatamente em uma só classe. Estas variações se podem ser explicadas por que no intervalo de 1 metro do testemunho descrito e analisado, podem ocorrem mais de uma litologia e zonas de transição entre elas. Concluímos que o desempenho dos dois modelos foram muito semelhantes , provando que o Random Forest funciona bem com muitas variáveis, ou que talvez ele apenas precisa das 4 variáveis escolhidas no modelo 1 para esses dados.

Foram realizados testes para a inclusão de profundidade como um atributo para o Random Forest. Porém, sua inclusão gera mais viés pois classes litológicas como R8 estão presentes em profundidades baixas onde o furo não é profundo e em profundidades altas até 40 metros com o espessamento do perfil. Também foram realizados testes com o algoritmo não supervisionado K-means para 4 elementos onde os *clusters* não corresponderam bem às classes.

O Random Forest consegue auxiliar na análise de dados dos bancos de dados com descrições de campo e laboratório afim de diminuir os erros atribuídos à essas atividades. O algoritmo conseguiu manter bem o balanceamento das classes para que não houvesse viés na realização das previsões.

Concluímos que o emprego do algoritmo Random Forest Classifier na classificação das amostas da base de dados do depósito de niquel laterítico do Rios dos Bois foi efetivo para reclassificar e entender melhor a relação dos dados e suas variáveis. Podendo ser possível aplicar em outras bases de dados para melhor entendimento e modelamento das mesmas.

Para o futuro espera-se continuar com o uso da ferramenta para publicações de artigos e melhor aplicação dos métodos.

6 REFERÊNCIAS

Barbour, A. P., 1976, Geologia do Maciço Ultramáfico de Santa Fé, Goiás. Tese, Instituto de Geociências, Universidade de São Paulo, p. 138.

Brand, N.W., Butt, C.R.M., and Elias, M., 1998, Nickel laterites: Classification and features. AGSO Journal of Australian Geology and Geophysics, v.17, p. 81–88.

Breiman, L., 1996. Bagging predictors. Mach Learn 24, 123-140.

Breiman, L., 2001. Random Forests. Statistics Department, University of California. Machine Learning, 45, 5–32.

Efron, B., Tibshirani, R., 1993. An introduction to the bootstrap. Chapman
& Hall, New York.

Elias, M., 2002. Nickel laterite deposits – geological overview, resources and exploitation. CODES Centre ofr Ore Deposit Research. Special Publication 4. 205-220.

Gleeson, S.A., Butt, C.R.M. and Elias, M., 2003, Nickel laterites: A review. SEG Newsletter, nº 54.

Golithly, J. P. 2010. Progress in Understanding the Evolution of Nickel Laterites. Society of Economic Geologists, Inc. Special Publications, v. 15, pp. 000–000.

Hoffman, P. 1997. Dna visual and analytic data mining. Visualization Conference, IEEE, 0:437.

Jooshaki, M.; Nad, A.; Michaux, S. A Systematic Review on the Application of Machine Learning in Exploiting Mineralogical Data in Mining and Mineral Industry. Minerals 2021, 11, 816. https:// doi.org/10.3390/min11080816

Kortchmar M. M. 2021. Geoprocessamento e machine learning aplicados à elaboração de mapa de favorabilidade para ocorrência de mineralizações Au-Pd-Pt na região de Serra Pelada, Província Carajás. Trabalho de conclusão de curso, Instituto de Geociências, Universidade Federal do Rio de Janeiro.

Machado, M. L. 2018. Distribuição geoquímica no regolito do depósito de níquel laterítico de Santa Fé, GO. Trabalho de conclusão de curso, Instituto de Geociências, Universidade Federal do Rio de Janeiro.

Martins, T. F. 2021. Mapas de potencial mineral dos sistemas mineralizantes de Cu-Au na folha Rio Verde, leste da Província Mineral de Carajás. Dissertação de Mestrado, Instituto de Geociências, Universidade Federal do Rio de Janeiro.

Oliveira, S.M.B., 1980, Alteração intempérica das rochas ultrabásicas de Santa Fé e gênese do depósito niquelífero. Tese, Instituto de Geociências, Universidade de São Paulo, p. 216

Pena, G. S. et al. Projeto Goiânia II, DNPM/CPRM, Goiânia, 1975, il., mapas e fotos, 236p.

Pimentel, M. M., Gioia S. M. L. C., Rodrigues, J. B. 1999. Geocronologia e geoquímica de ortognaisses da região entre Iporá e Firminópolis: Implicações para a evolução do Arco Magmático de Goiás. Rev. Bras. Geoc., 29(2):207-216.

Putzolu, F. et al. 2020. The Influence of the Magmatic to Postmagmatic Evolution of the Parent Rock on the Co Deportment in Lateritic Systems: The Example of the Santa Fé Ni-Co Deposit (Brazil). Economic Geology.

Salzberg, S.L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach Learn* 16, 235–240 (1994). https://doi.org/10.1007/BF00993309

Zhou, Z. 2021. Machine Learning. Nanjing University, China. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, Springer.

Orange Data Mining (https://orangedatamining.com).

Folha geológica Iporá 1:100000 CPRM (https://geosgb.cprm.gov.br).