INICIAÇÃO CIENTÍFICA
PIBIC-USP

INTELIGÊNCIA ARTIFICIAL, *MACHINE LEARNING* E *DEEP LEARNING* EM EXPLORAÇÃO MINERAL E GEOLOGIA ECONÔMICA: USOS E APLICAÇÕES

Gabrel Góes Rocha de Lima

Orientador: Prof. Dr. Caetano Juliani

RELATÓRIO PARCIAL

Março de 2021

RESUMO

Com o aumento da capacidade de computação das máquinas, redução de custos de *hardwares* e o contínuo desenvolvimento de algoritmos na área da inteligência artificial (IA), as técnicas de aprendizagem de máquina estão sendo cada vez mais aplicadas e nos mais diversos ramos das ciências e da indústria. Nas ciências da terra, as técnicas de IA tem uma enorme gama de aplicações nas mais diversas áreas e, atualmente ocorre um grande desenvolvimento aplicado à busca de novos depósitos minerais, área esta conhecida como prospecção ou exploração mineral. A IA opera por meio de algoritmos, que são passos computacionais a serem seguidos pela máquina, capazes de reconhecer padrões e classificar dados de forma semi-automática ou automática. Os principais algoritmos de aprendizagem de máquina aplicados nas geociências, com foco na exploração mineral, são o *Support Vector Machine*, *Random Forests*, *Artificial Neural Networks*, *Naive Bayes* e *k-Nearest Neighbors*, entre outros. Estas técnicas são capazes de classificar quantidades enormes de dados geoespaciais em diferentes escalas, desde imageamento de lâminas petrográficas até imagens orbitais, passando por perfilagem (*logging*) e levantamentos aeroportados (Aeronaves ou VANTs).

Esta iniciação científica visa a revisão da literatura sobre o tema de *machine learning* aplicada à mineração, buscando o entendimento de suas técnicas, funcionalidades e aplicabilidades aos empreendimentos brasileiros de pesquisa e exploração mineral, construindo um arcabouço teórico do estado da arte, seguida da produção de um material metodológico didático no formato *Jupyter Notebook* como um guia dos métodos aplicados e disponibilizado na plataforma Github*.

SUMÁRIO

1 INTRODUÇÃO

Nas últimas décadas, a produção e análise de dados geólogicos intensificou se devido às maiores facilidades de coleta de amostras e facilidades e custos analíticos, além da ampliação da bsca por novos depósitos minerais. Esse enorme volume de dados exige processamentos estatísticos mais complexos e sistemas de armazenamento acessíveis, robustos e integrados a em vários sistemas. O grande volume de dados torna praticamente impossível a análise convencional em algumas áreas, com dezenas de milhares dados analíticos variados, incluindo aerogefisica, geoquímicos de rochas, sedimentos de corrente, solos, minerais,mineralógicos, termobarométricos, etc, razão pela qual o uso de técnicas de IA tem se tornado necessárias no tratamento dos dados e na busca de depósitos minerais. Além disto, no caso específico de muitas áreas no Brasil, a escala dos mapeamentos geológicos ainda não é a adequada à exploração mineral e as coberturas de solo e de florestas e a falta de acessos dificultam a obtenção de dados e a definição do potencial geológico para ocorrência de depósitos minerais. Neste contexto, os métodos propostos neste trabalho de Iniciação Científica, podem ser importantes indicadores de potencial onde faltam informações geológicas.

Desta forma, com o auxílio das tecnologias recentes e com a quantidade de dados aerogeofícios disponibilizados pelo Serviço Geológico Brasileiro (SGB), é possível gerar modelos de exploração mineral por meio de IA, porém, para se construir modelos preditivos de forma eficiente é necessária a construção de uma base de dados organizada e robusta que possibilite a fácil manipulação, seja para a extração de dados para a produção de modelos, seja para inserção de novos dados gerados.

Este projeto de pesquisa foca o estudo, baseado no case do bem mineral grafita, do potencial de aplicação de técnicas de *machine learning* na exploração mineral. Como as aplicações em geociências são relativamente recentes, deverá ser também elaborado material metodológico didático.

2 OBJETIVO

O objetivo desta pesquisa é o treinamento e aprendizado dos conceitos e técnicas de Inteligência Computacional e de suas potenciais aplicabilidades à

geologia econômica e exploração mineral, possibilitando a produção de um mapeamento litológico preditivo que auxiliará a prospecção mineral, além da produção de um conteúdo didático dos métodos estudados e aplicados nesta pesquisa.

Trata-se, portanto, de um projeto de pesquisa que se inicia como uma revisão bibliográfica do estado da arte das mais variadas técnicas de Inteligência Computacional, seguida da escolha de um método a ser aplicado é descrito como apostilas didáticas no formato *Jupyter Notebook* disponibilizada na plataforma GitHub.

3 CONTEXTUALIZAÇÃO DO PROBLEMA CIENTÍFICO

Depósitos minerais são anomalias geoquímicas na crosta terrestre, ou seja, a ocorrência de concentrações anômalas de um determinado elemento químico (ou associações) são de ocorrência local e raras na crosta terrestre.

A demanda por bens minerais aumenta continuamente, principalmente como resultado do aumento da população mundial, incremento da população em cidades, desenvolvimento de infraestrutura e maior acesso aos bens de consumo. Em adição, o desenvolvimento tecnológico tem propiciado inúmeros bens de consumo novos (como computadores e celulares) e novas tecnologias mais sustentáveis (como a energia eólica e veículos elétricos), que têm utilizado um número crescente de elementos químicos relativamente os produtos comuns e, consequentemente, de novos minérios.

A descoberta de depósitos minerais tem exigido altos investimentos, com relativamente pouco retorno, pois os depósitos têm sido encontrados em níveis crustais cada vez mais profundos e em regiões ínvias. Por estes motivos, novos conceitos exploratórios, novos modelos de gênese de concentrações de minérios e novas técnicas exploratórias, como as aqui propostas, têm sido necessárias.

A busca de soluções para este conjunto de problemas tem sido cada vez mais focada no uso de inteligência artificial, com aplicação de algoritmos e técnicas diversas que permitem a descoberta de padrões e correlações de variáveis em sistemas naturais que indicam, com maior nível de segurança e menor custo, por redução das áreas da pesquisa, de custos analíticos e de exploração indireta, de

locais mais potenciais para ocorrência de depósitos minerais.

Este projeto insere-se neste contexto, visando a formação complementar do bolsista nesta importante área das geociências voltada para descoberta e produção de recursos minerais essenciais para o desenvolvimento socioeconômico do país e da sua população.

4 JUSTIFICATIVAS PARA A PESQUISA

Justifica este trabalho o treinamento do aluno no uso de ferramentos de IA em geociências, com foco na exploração mineral e na geologia econômica, e a produção de materiais educacionais que poderão ser utilizados por outros alunos do curso de Geologia do Instituto de Geociências da USP.

A escolha de uma área-modelo para o desenvolvimento destes estudos foi feita com foco em novos materiais de usos tecnológicos, tendo sido escolhida a grafita. Baseado nas pesquisas bibliográficas, foi escolhida a região nordeste do estado de São Paulo e sudoeste de Minas Gerais devido à ocorrência de corpos de grafita xistos de alto grau metamórfico, ainda não estudados ou explotados. A grafita de alto grau de cristalinidade é necessária para uma série de aplicações tecnológicas, destacando-se a produção de baterias utilizadas em veículos elétricos (EVs). Este tipo de grafita é relativamente raro e a região da serra da Mantiqueira em Minas Gerais e outras áreas na Bahia tem mostrado potencial para ocorrências deste tipo de depósito mineral. Os mapeamentos geológicos na região em estudo evidenciam também corpos de rochas grafitosas metamorfisados na fácies granulito, de modo semelhante às demais regiões potenciais citadas, justificando esta proposição.

Usualmente os corpos de rochas grafitosas mostram-se relativamente às rochas encaixante, enriquecidos em urânio e sulfetos e estes elementos podem gerar assinaturas que podem ser detectadas por meios indiretos, como imagens de satélite (dado a alterações na cobertura vegetal) ou em imagens de aerogeofísica,

dados os comportamentos do U-Th-K. Estes dados, juntamente com as informações geológicas disponíveis, estão sendo avaliadas por IA, visando, em primeira instância, a delimitação dos corpos de rochas grafitosas, por meio de modelos preditivos. Após esta fase, serão buscados informações que possam indicar maior potencial para ocorrência de de corpos com dimensões, teores ou qualidade de importância econômica, se possível, dada a pandemia, com visita em campo para amostragens.

5 RESULTADOS PARCIAIS

5.1 Pesquisa Bibliográfica

A pesquisa bibliográfica do projeto tem sido feita nos bancos de dados Scopus, Scielo, SibiUSP e Google Acadêmico com as palavras chaves *Machine learning/Deep Learning/Random Forests*. As principais revistas revisadas por pares encontradas foram *Computers & Geosciences*, *Ore Geology Reviews* e *Journal of Coal Geology*, de onde foram retirados a maioria dos artigos no presente trabalho. Além dos cursos online, práticos e teóricos, que possibilitaram o início da etapa de aplicação das técnicas e a produção das apostilas didáticas.

5.2 Bases de Dados Aerogeofísicos e Geológicos

O Brasil possui uma área enorme de cobertura de levantamentos geofísicos e geológicos (Fig. 1) catalogada sistematicamente de forma padronizada, o quê abre uma enorme gama de tratamentos possíveis com foco na exploração mineral e, até mesmo, para subsidiar mapeamentos geológicos e outras aplicações relacionadas às geociências. Uma base de dados que permita acessar e armazenar a informação de forma rápida e eficiente é essencial para a construção de modelos preditivos em grande escala e automatizados.

Com base neste dados geológicos foram aplicados algoritmos de aprendizagem de máquina aplicados nas geociências, com foco na exploração mineral, são o *Support Vector Machine*, *Random Forests*, *Artificial Neural Networks*, *Naive Bayes* e *k-Nearest Neighbors*, entre outros.

Com a coleta e avaliação destas informações geoespaciais foi possível identificar as áreas onde há levantamentos geológicos e geofísicos em escalas compatíveis que auxiliarão a produção dos dados de ocorrências minerais, assim

como a correlação entre os diferentes terrenos geológicos e suas ocorrências minerais.

Para esta pesquisa foi feito o *download* de toda a base de metadados dos projetos de aerogefísica, dos mapas de ocorrência mineral e geológicos

disponibilizados pelo SBG. Em, seguida da correção de erros de armazenação da informação, como nos vetores ‘Projetos de Aerogefísica’ em que a escolha incorreta das classes dos valores de distâncias entre linhas de voo da coluna ‘ESPACAMENT” salva como uma ‘*string*’, quando deveriam ser salvas como um “*integer*” ou “*float*”, remoção de dados expúrios, onde haviam alguns levantamentos com ‘ESPACAMENT’ incoerentes como distâncias entre as linhas de voo marcadas em escalas unitárias, provavelmente em quilômetros, enquanto a maior foi descrita em metros.

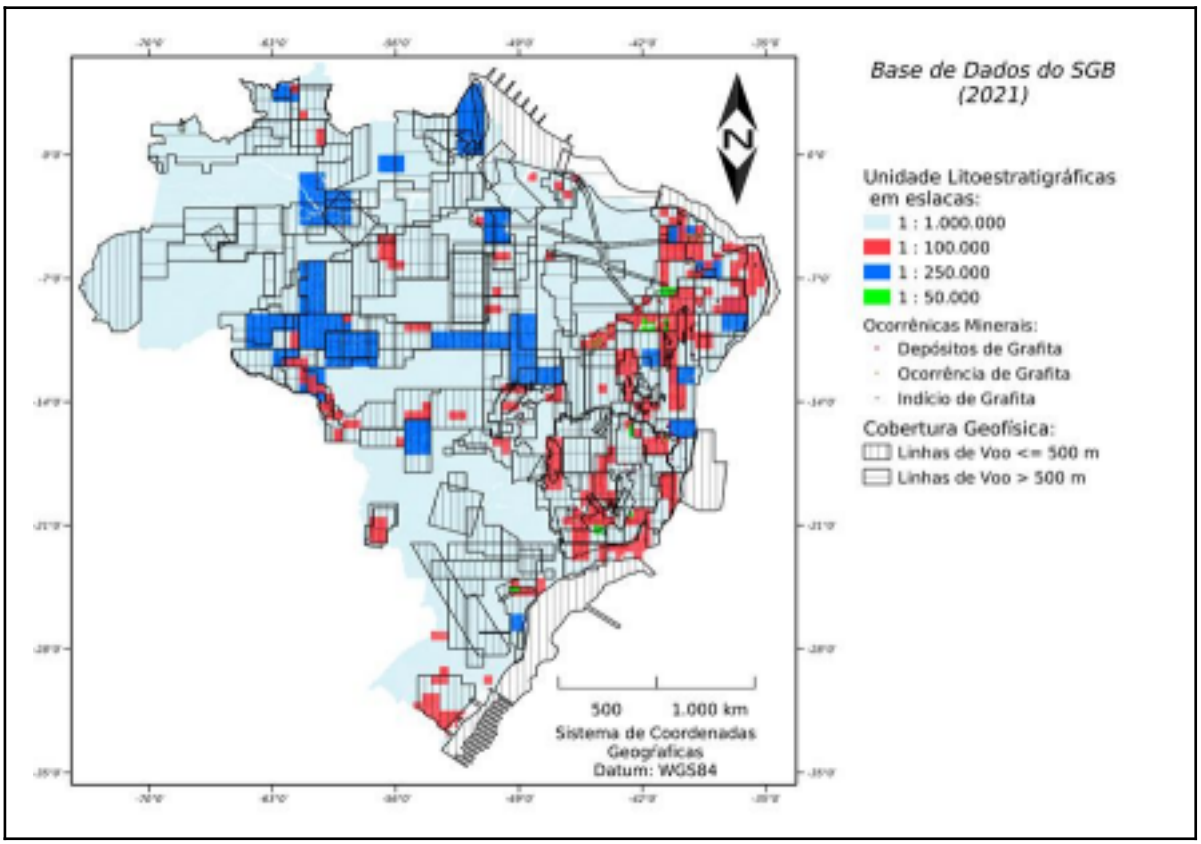

Figura 1 *– Mapa-índice dos levantamentos aerogeofísicos disponibilizados pelo SGB.*

Com a sobreposição dos mapeamentos geológicos e ocorrências minerais pelos projetos aerogeofísicos disponibilizados pelo SGB, foi possível identificar áreas promissoras para a aplicações das técnicas de aprendizado de máquina. Com isso,

foi selecionada uma área com dados suficientes para a aplicação das técnicas de IA.

A análise dos mapas e relatórios geológicos indica que as áreas com mineralizações de grafita estão relacionadas a eventos de metamorfismo colisional orogênico pré-cambrianos e, portanto, se encontram nas margens dos terrenos cratônicos da borda leste da Plataforma Sul Americana (Fig. 2).

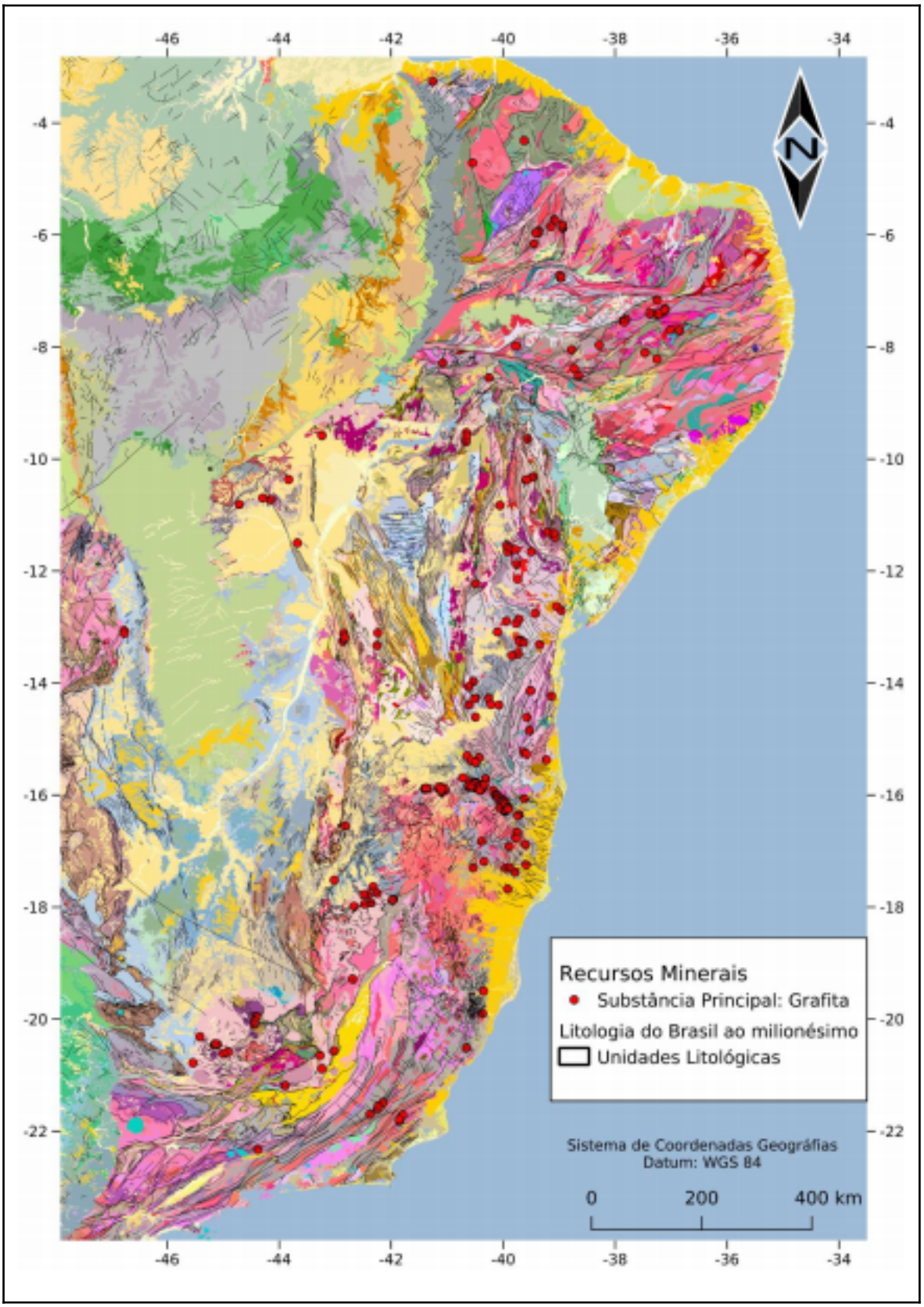

Figura 2 *–Recorte da porção litorânea do mapa geológico do Brasil ao milionésimo, com as ocorrências de grafita.*

Duas áreas com concentrações de ocorrências de grafita foram selecionadas (Fig. 3 e 4). A existência de mapas geológicos de maior detalhe [1, 2, 5] com ocorrências de grafita não destacadas nos levantamentos do SGB [3, 4, 7, 12] conduziu à seleção desta área para os estudos, por possibilitar a indicação de corpos mineralizados em grafita e, sobretudo, suas extensões em mapa, indicativo de potencial para exploração mineral, juntamente com dois fatores geológicos importantes para geração de depósitos de grafita de alta qualidade tecnológica, que são o grau metamórfico alto, a existência de zonas de cisalhamento em regimes térmicos elevados e com alterações hidrotermais [1, 5], inclusive, com ocorrências de ouro associadas.

5.3 Avaliação das Informações Geológicas e Aerogeofísicas

Após a fase inicial de estudo dos conceitos e técnicas computacionais aplicadas às geociências [3, 6, 7, 8, 10, 11, 13], foi iniciada a etapa de compilação e avaliação das informações geológicas, visto que a aplicabilidade destes conceitos e técnicas depende do *commodity* mineral, do tipo do depósito, da geologia regional e da disponibilidade de dados.

Com base nos levantamentos bibliográficos foram inciados os processamentos das informações coletadas. Ecom isso foi possível identificar as características geológicas necessárias para possibilitar a geração de modelos preditivos confiáveis, orientando a definição de áreas potenciais para exploração de grafita. Para tanto foram utilizados mapeamentos geológicos em diferentes escalas e, em especial os de maior detalhe [1, 2 , 5], cujas informações foram correlacionadas com os diferentes levantamentos disponibilizados pelo Serviço Geológico do Brasil.

Com a continuidade da pesquisa serão vetorizados os mapas geológicos das folha Socorro (Fig. 5) e Campinas (Fig. 6), cujas escalas possibilitarão melhorias na classificação, pois ambos os mapas estão em escala mais compatível com a escala do levantamento aerogeofísico de código “1039” (Fig. 1).

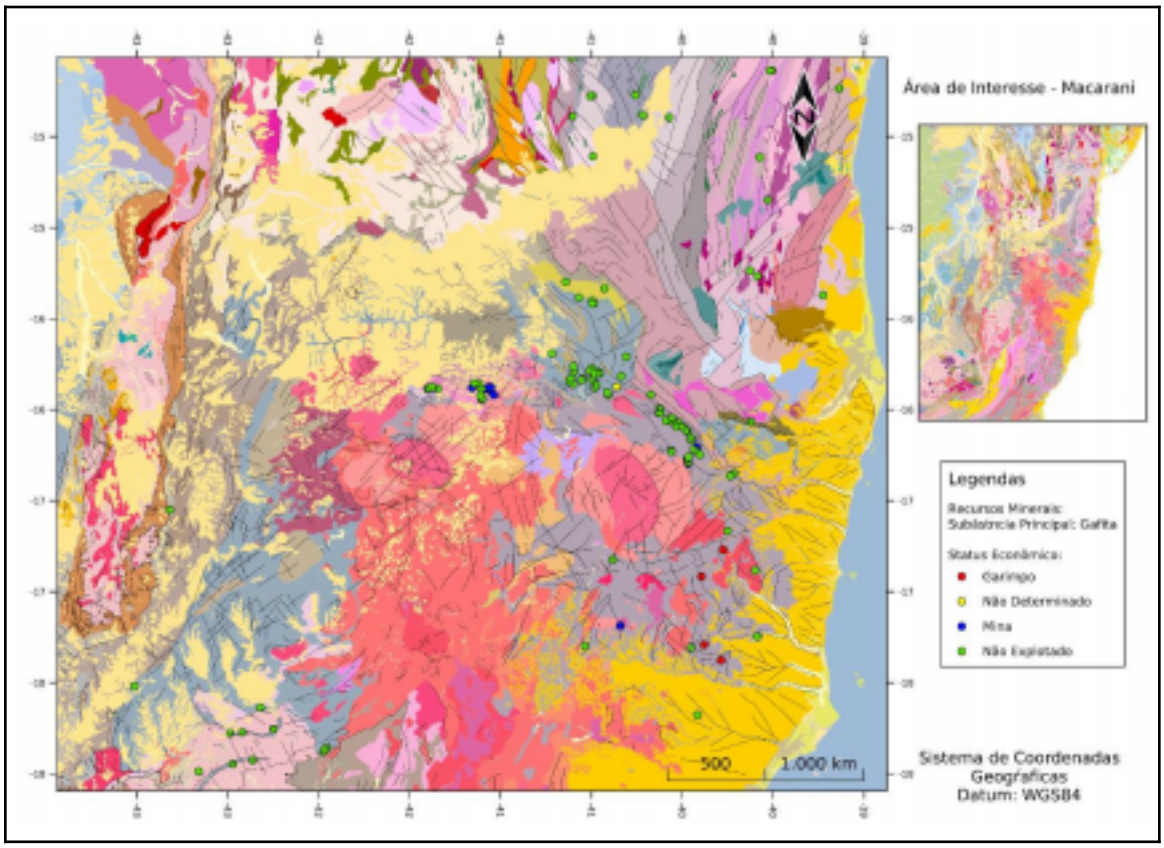

Figura 3 – *Detalhe do mapa anterior, com a área de interesse Macarani.*

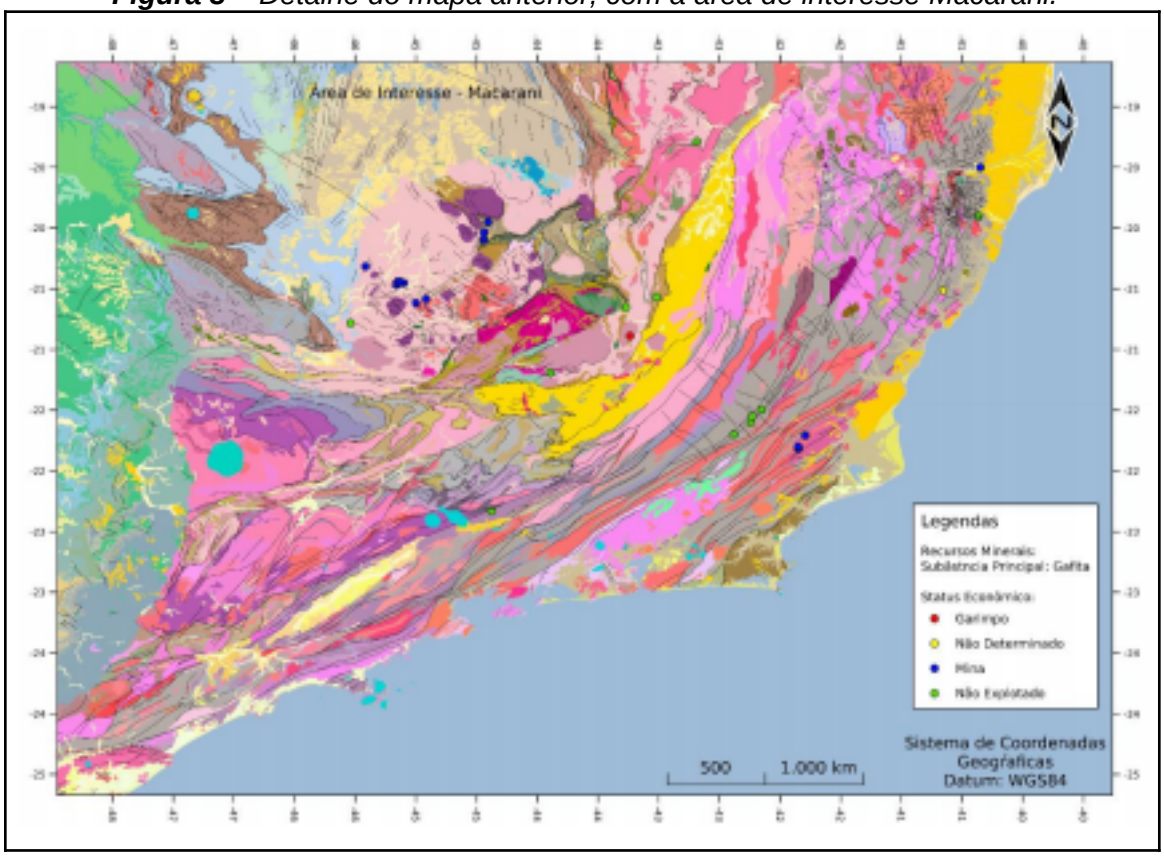

Figura 4 – *Detalhe do mapa a Figura 2, com a segunda área de Interesse, situada nas proximidades da divisa de São Paulo e Minas Gerais.*

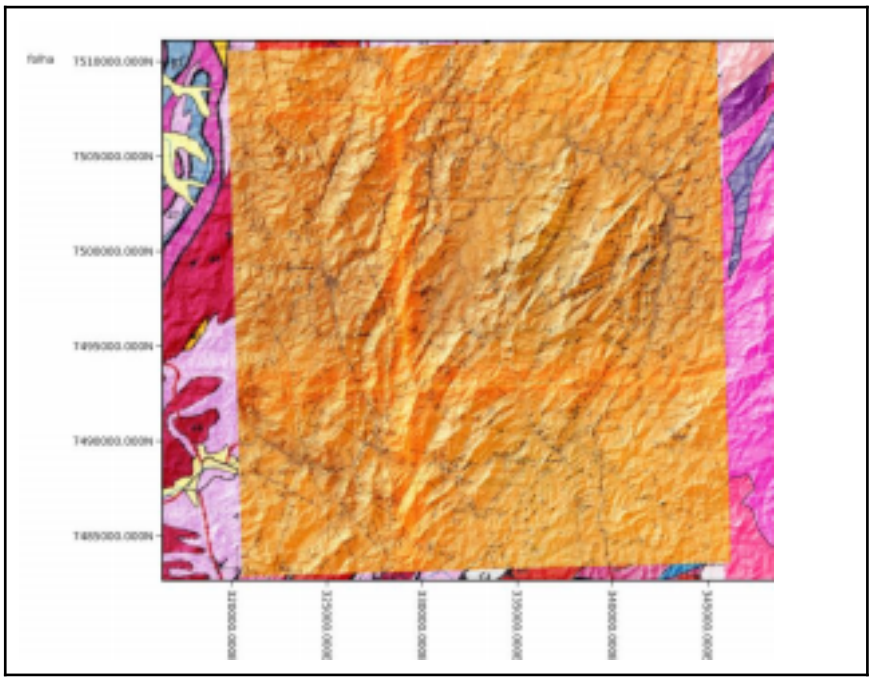

Figura 5 *– Recorte do mapa geológico da Folha Socorro em 1:50.000 com a sobreposição do modelo digital de terreno sombreado gerado a partir das imagens ALOS/PALSAR disponibilizadas pela Agência Espacial Japonesa (JAXA), onde há ocorrências de rochas grafitosas.*

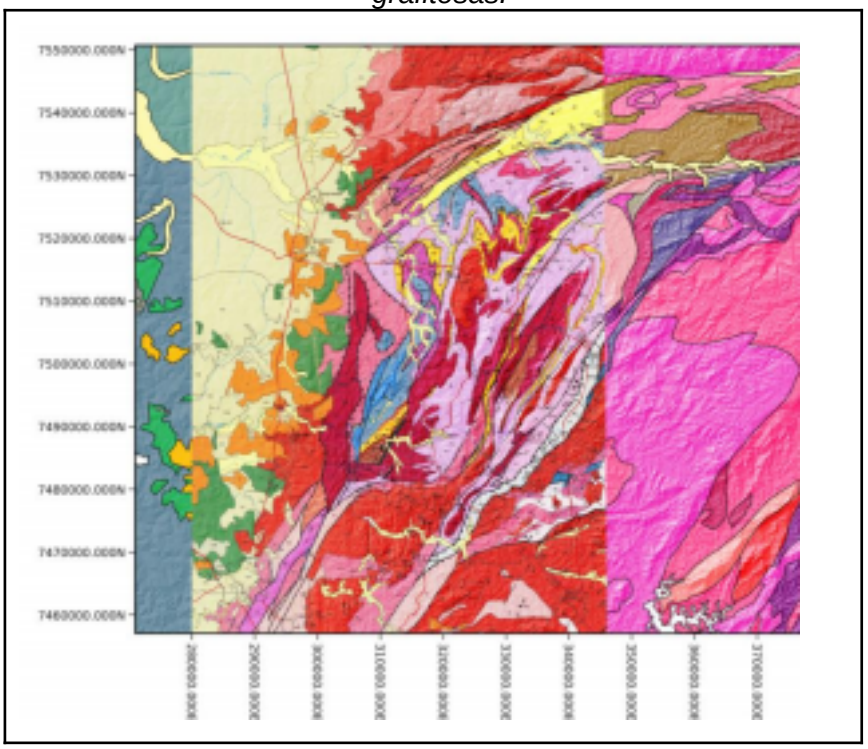

Figura 6 *– Recorte do mapa geológico da Folha Campinas em 1:250.000, sobreposto ao mapa geológico do Brasil ao Milionésimo (Folha SF-23), com modelo digital de terreno sombreado sobreposto, da mesma fonte da figura anterior.*

O mapeamento de [5] estão em escala muito maior (1:25.000), não sendo possível correlações com com os dados dos levantamentos aerogeofísicos pois, as dimensões das unidades litotípicas mapeadas, em especial daquelas que contém as rochas grafitosas, são muito menores do que as distâncias entre as linhas de voo percorridas pelos aviões, o que resulta em pixels híbridos, subamostragem das litologias menores que as da linha de voo, dificultando a identificação de padrões e correlações, representada pela diminuição do score das predições da banda de urânio geradas pela função da biblioteca de códigos livres, *Verde,* de *vd.CrossValidation*. Entretanto, corpos mineralizados de interesse econômico podem ocorrer na escala deste mapa e a análise deste potencial, mesmmo com pixels híbridos, pode ser importante para exploração mineral.

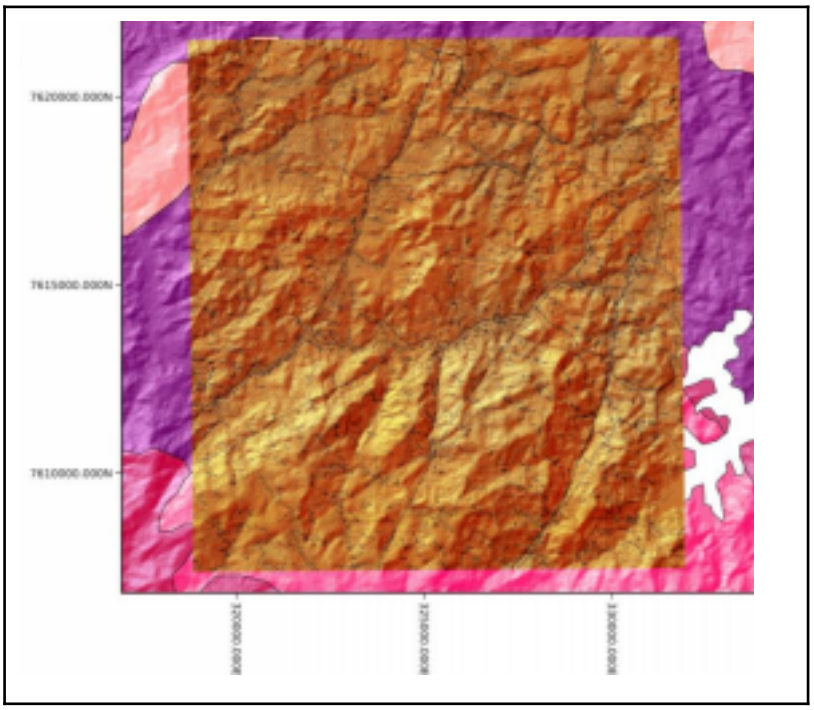

Figura 7 *– Recorte do mapa geológico do quadrante noroeste da Folha Caconde em 1:50.000 com modelo digital de terreno sombreado gerado a partir das imagens ALOS/PALSAR disponibilizadas pela Agência Espacial Japonesa (JAXA) sobreposto.*

Os mapas geológicos das Figuras 3 e 4 estão em escalas mais compatíveis com a do levantamento aerogeofísico 1105. Estes mapas já estão sendo vetorizados para posteiror início do *workflow* no *software Orange*.

Estes mapas geológicos serão muito importantes pois, além de apresentarem melhor resolução das estruturas geológicas e litologias, possuem também dados

geoquímicos de sedimento de corrente, dados de sondagens e de descrição de

amostras de lâminas petrográficas, os quais poderão ser muito úteis na elaboração dos modelos exploratórios por IA.

Três levantamentos aerogeofísicos foram feitos na área de estudo, identificados pelos os códigos de ‘1039’, ‘1105’ e ‘Área14’. Os metadados destes levantamentos foram obritos no *site* RIGEO da CPRM (Serviço Geolóogico do Brasil). O levantamento ‘1039’ foi feito com linhas de voo espaçadas em 1000 m e os demais, mais recentes, em linhas de voo de 500 m. Estes também possuem maior qualidade nas amostragens, sensores e no pré-processamento dos dados.

O levantamento aerogeofísico se da pela coleta de dados por um sensor aeroportado de contagens de radiações gama e de variação do campo magnético. Esta radiação vem da camada superficial das rochas da área sobrevoada, e possui uma variação no comportamento energético, variância que representa a origem/fonte da radiação, sendo elas classificadas em bandas energéticas, cada uma delas representando uma fonte de radiação específica, sendo elas as bandas de contagens de Tório (Th), Urânio (U) e o Potássio (K).

A partir destas contagens, e do tratamento dos dados coletados pelos sensores, como a correção das contagens totais a partir da contagem de radiações de fundo coletadas por um outro sensor apontado no sentido oposto, é possível manipular matematicamente os valores, como a concentração de contagens de radiações provenientes decaimento do potássio (K-40) pelas contagens totais (S) coletadas pelo sensor, produzindo uma banda nova gerada a partir de duas outras variáveis, criando correlações importantes para o método de classificação automática por algoritmos de aprendizagem de máquina.

O Fator F é gerado pela manipulação matemática dos valores coletados em cada banda, que tem uma grande importância na separação das classes litológicas [4, 9 e 10].

5.4 Estudo dos Métodos de Tratamento dos Dados e Algoritmos de Classifcação Automática

Após a etapa de seleção e reconhecimento da área de estudo, foi dado início ao estudo e processamentos dos métodos de interpolação e de algoritmos de classificação, possibilitando a criação do fluxo de trabalho (*workflow)* desde os

códigos em Python para tratamento dos dados brutos dos levantamentos aerogeofísicos até a interpolação e a contrução do fluxograma do algorítmo de apredizagem de máquina. Estes estudos focaram também a organização e validação dos dados produzidos pela sistematização das técnicas aprendidas.

5.5 Aplicação do Método de Splines

A nuvem de pontos geradas pelos levantamentos aerogeofísicos não é regularmente distribuída, sendo alterada pelos erros nos aparelhos GPSs condicionados à quantidade de bases terrestres (*ground stations)*, além de ser amostrada com frequência muita maior ao longo das linhas de voo quando comparada com as distâncias entre as linha. Esta irregularidade espacial na amostragem de dados impede que muitos métodos de processamento de dados sejam concluídos com êxito, como o erro de *aliasing effect,* produzido quando se interpola uma dado amostrado com maior frequência em uma direção do que em outra, gerando linhas retas e paralelas ao longo do eixo super amostrado, confundindo a interpretação, caso se deseje reconhecer lineamentos nas imagens [13].

Com a técnica de interpolação pela *Green's function*, é possível corrigir tanto o problema da irregularidade espacial da amostragem, quanto o problema das grandes distâncias entre pontos de amostragem em um levantamento geofísico aeroportado. O modelo de regressão linear é gerado a partir de uma função que descreve o comportamento dos valores entre os valores amostrados, em que uma variável descreve o quão acentuada serão as curvas, que seria o grau da função polinomial a ser gerada.

Estudos mais aprofundados sobre esta função estão sendo feitos neste momento da pesquisa, pois, a atribuição desta variável que descreve o quão acentuada são as curvas não pode ser arbitrária e caso seja feita de forma ótima, podemos resolver os problemas encontrados nas regressões lineares das bandas que apresentam uma alta variância dos dados amostrados, como as bandas de urânio. O resultado parcial para a variável urânio pode ser visto na Figura 8.

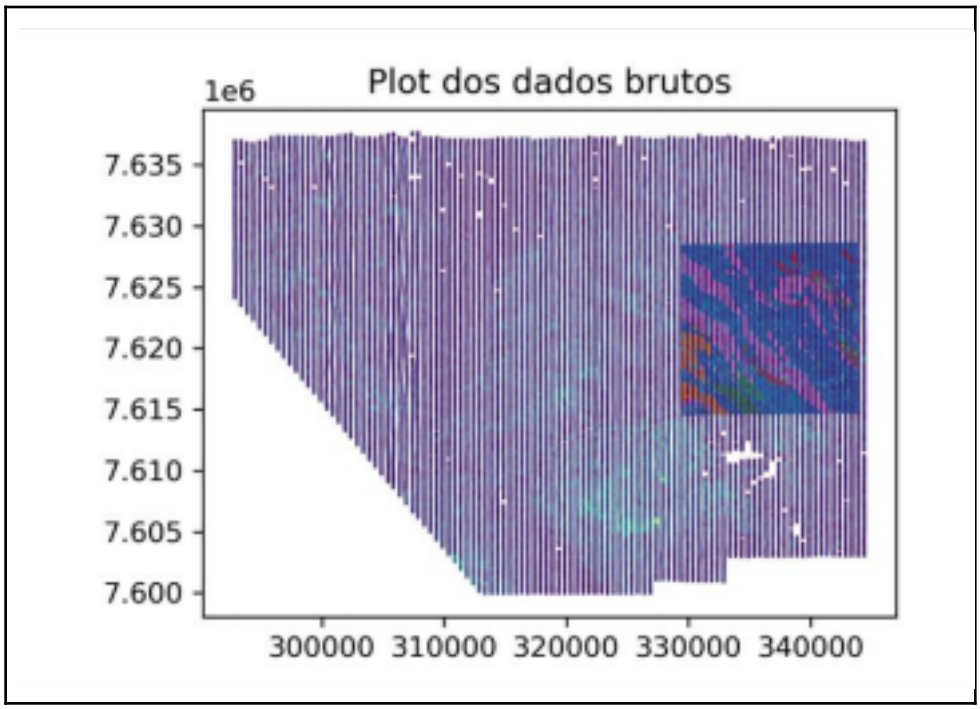

Figura 8 *– Plot dos dados de urânio equivalente brutos do levantamento de código '1105' sobreposto pelo mapa geológico de [5].*

Esta função é apresentada na Figura 9. A função recebe todos os pontos amostrados pelo aerolevantamento e os reduz para uma amostragem com valores à uma distância média de 500 m. O procedimento reduziu o número de pontos de 43.413 para uma malha de 5.964 pontos, como mostra a Figura 10**.** Este passo é essencial para que as operações algébricas possam ser realizadas e para que o processo de interpolação não gere *grids* com efeitos de *Alliasing* provocados pela maior amostragem dos dados no eixo Y (frequência de captura do sensor X velocidade de vôo da aeronave) do que no eixo X (distância entre as linhas de voo).

```
# ------------------------- BLOCKED REDUCTIONS ------------------------------
# 1 - vd.BlockReduce---------------- CREATING A REDUCER  ---------------------
# Block reduction are dividing the region in blocks of a especified spacing
# -- Setting Coordinates
coords_1105 = (dados['UTME'].values, dados['UTMN'].values)

# -- Creating a reduction function with 'np.median'
reducer = vd.BlockReduce(np.median, spacing=500)
# -- Reducing the data by sampling points at a median distance of 1000 m.
b_coords, b_eU = reducer.filter(coords_1105, dados.eU)
```

Figura 9 *– Linha de comandos em Python3 da função verde BlockReduce no sofware Orange.*

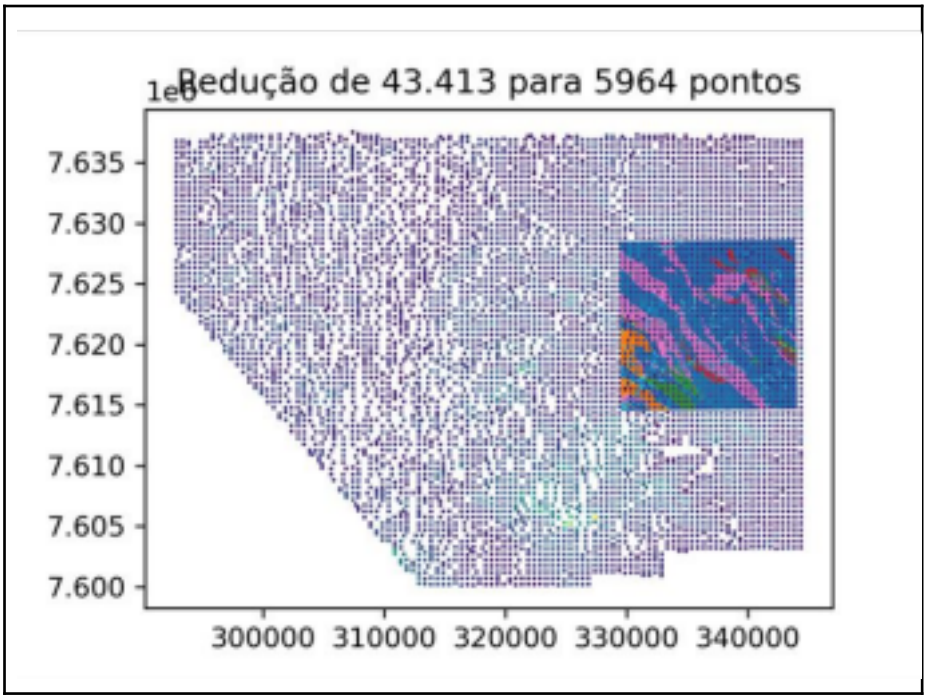

Figura 10 *– Plot dos dados reduzidos à espaçamentos de 500 m.*

```
# 2 - ---- FITTING THE LINEAR MODEL OF DECIMATEDGRID WITH SPLINE  ------------
# ------------------------- Selecting the Linear model as a vd.spline
spline = vd.Spline()
# def spline.fit():
spline.fit(b_coords, b_eU)


# 3 -------- PREDICTTING THE ACTUAL DATA WITH THE LINEAR MODEL ---------------
# the values of non-decimated dataset with the fitted linnear model
predicted = spline.predict(coords_1105)
```

Figura 11*– Linhas de comandos que executam a operação de "fiiting' da curva linear que descreverá o coportamento dos valores ao longo da variação espacial (#2) e comandos que realizam a predição dos dados nas coordenadas amostradas no levantamento.*

A Figura 12 representa a qualidade do modelo é na predição de valores em pontos amostrados, não representando, exatamente, o quão boa será a interpolação em um *grid* regular sintético. Portanto, foi gerado um *grid* sintético para que a função calcule os valores desta nova amostragem em um *grid* sintético regular (Fig. 13 e 14).

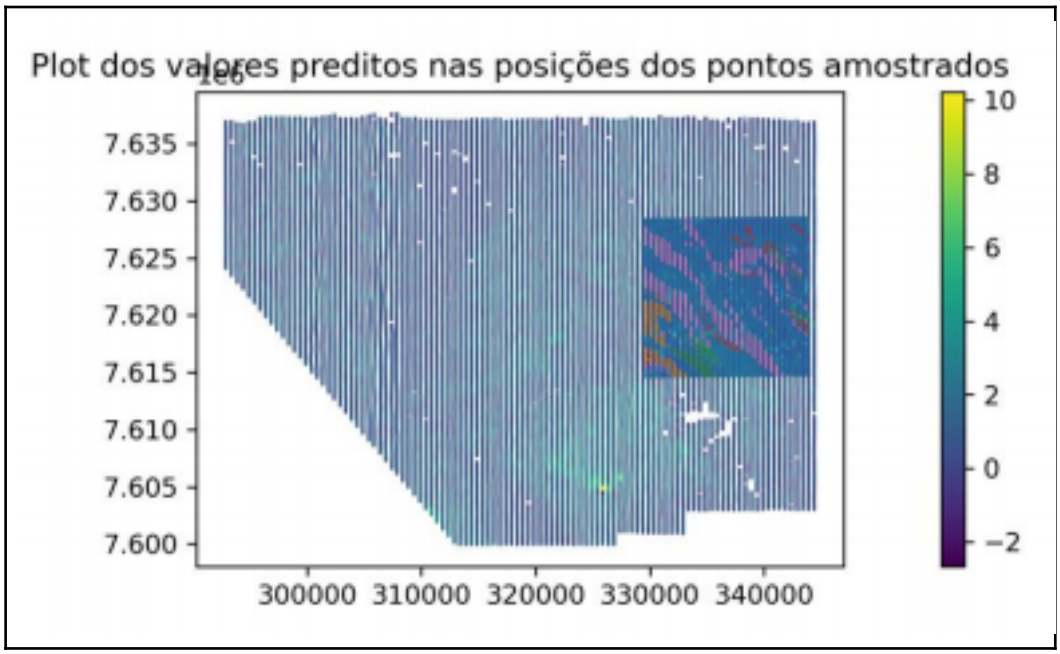

Figura 12 *– Plot dos valores preditos para posições amostradas.*

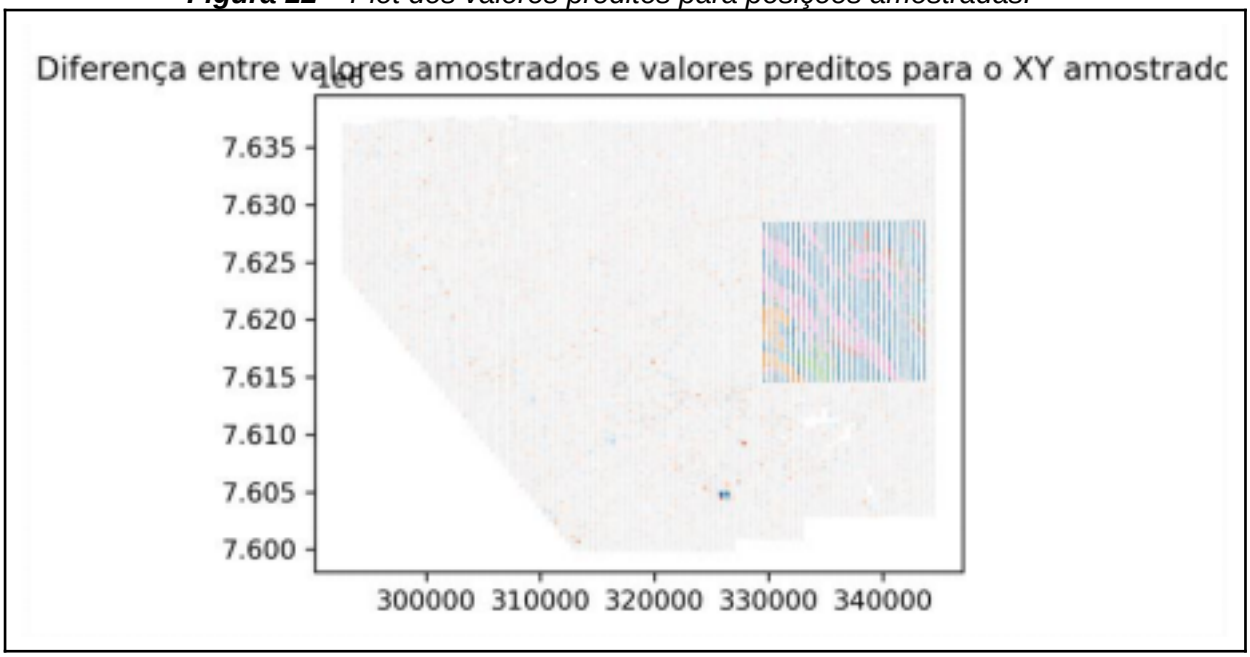

Figura 13 *– Plot da diferença entre os dados preditos para os dados amostrados.*

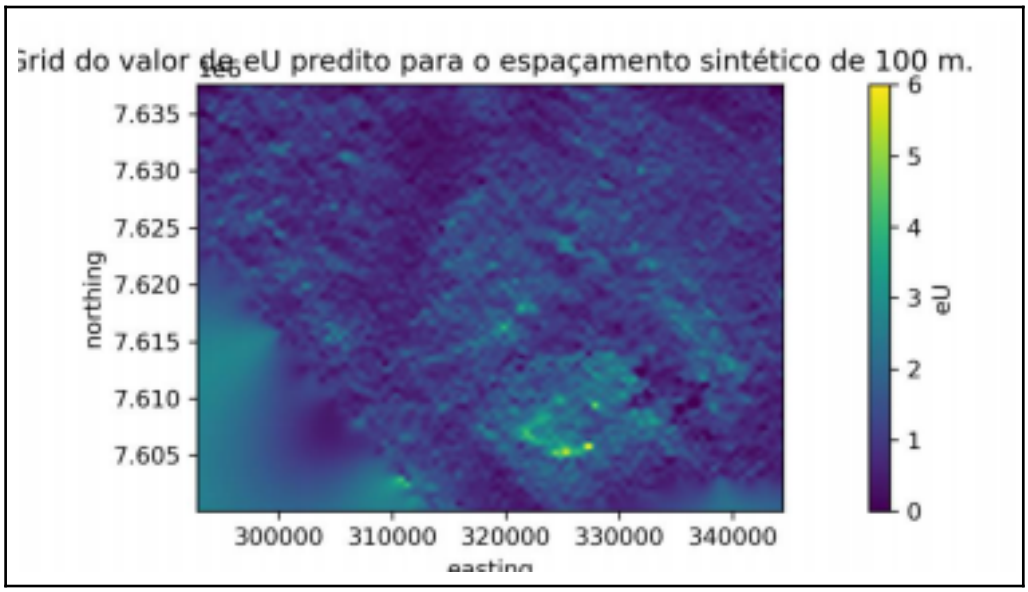

Figura 14 *– Plot do grid sintético espaçado em 100 m.*

5.6 *Outputs* da Etapa de Pré-Processamento

Adiante são apresentados os *outputs* (Fig. 14 e 15) da etapa de pré processamento dos dados dos levantamento aerogeofísicos do Projeto São Paulo - Rio de Janeiro realizado pelo convênio DNPM/CPRM de 1988. As informações que serão analisadas e concatenadas em um *GeoDataFrame* que posteriormente serão introduzidas no *software Orange* para serem classificadas com o algoritmo de *Random Forests*.

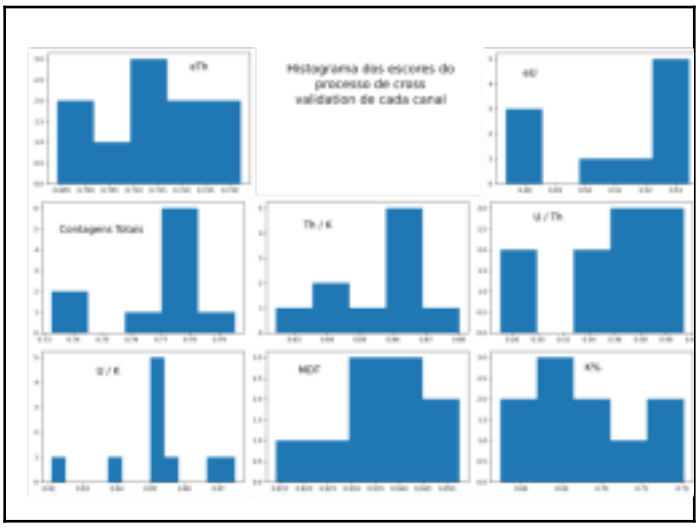

Figura 14 *– Histogramas dos escores do processo de cross validation de cada um dos canais do levantamento aerogeofísico onde: eTh = Tório equivalente, eU = Urânio equivalente, K% = Potássio em porcentagem, MDT = modelo digital de terreno, Th/K = razão Tório/Potássio, U/Th = razão Urânio/Tório, U/K = razão Urânio/Potássio e as*

contagens radiométricas totais.

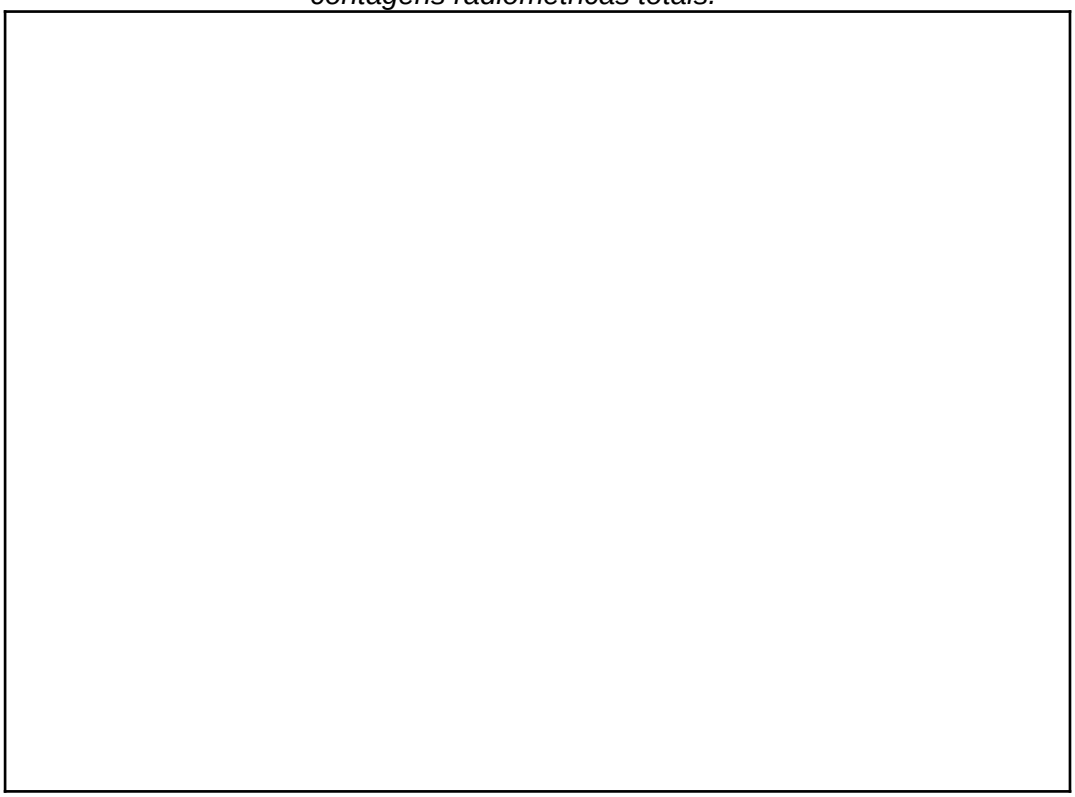

Figura 15 – *Representação da variação dos dados amostrados entre levantamentos adjacentes. As variáveis são as mesmas da figura anterior.*

Analisando os *plots*, tanto dos valores *gridados* quanto dos histogramas de escores do processo de *cross validation,* é possível identificar quais as bandas que poderiam gerar *features* com melhores *gini index* (variável que quantifica a competência de uma *feature* em separar um conjunto de dados em classes distintas). Com isto, pode ser reconhecido que a banda de contagens de urânio não resultou em bons *scores* no levantamento aerogeofísico com linhas de voo com espaçamentos de 500 m), assim como no levantamento com linhas espaçadas em 1000 m.

Outra observação possível é que a *feature* produzida pela razão de urânio/potássio tem boa acurácia, porém baixa precisão, enquanto que a razão urânio/tório apresentam baixa precisão e baixa acurácia.

É preciso salientar que as escalas mostradas nos *plots*, não representam o intervalo real dos valores encontrados, pois, por exemplo, apresentam intervalos negativos, apesar dos dados não chegarem àqueles valores.

Os estudos feitos até o momento permitem também identificar que as medidas de contagens gamaespectrométricas apresentam discrepâncias entre os dois

aerolevantamentos, sendo portanto necessários um pré-processamento para nivelamento dos dados. Esta diferença de medidas é exemplificada pelos plots das imagens interpoladas dos levantamentos aerogeofísicos de código “1105” e um terceiro levantamento, também amostrado à 500 m de código “Área 14”.

5.7 Método *Random Forests*

Etapa de montagem do fluxo de trabalho (Fig. 16) para aplicação dos algoritmos de aprendizagem de máquina no *software Orange*. Após a etapa de preprocessamento os dados gerados pela interpolação foram transformados de imagens (*.netCDF*), para um tabela (*.csv).*

Figura 16 – *Fluxograma (workflow) dos passos computacionais de aprendizagem de máquina executados na pesquisa.*

Na Figura 17 é exemplificado o modo e os valores calculados dos dados aerogeofísicos e de MDT para entrada no processamento *Random Forests* para treinamento do modelo.

Figura 17 – *Formato e exemplo de valores em que os dados aerogeofísicos são inseridos na etapa de treino do modelo.*

O histograma da litologia que utilizada como *target* do algoritmo, será a variável da classe litológica predita na etapa de teste (Fig. 18).

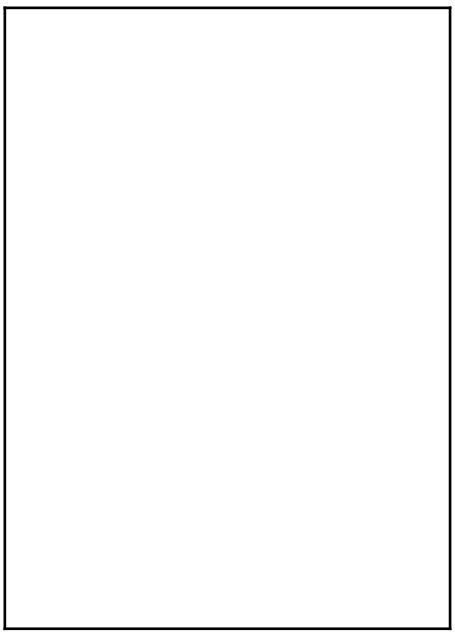

Figura 17 – *Histograma da distribuição das categorias dos dados "litologia" para a etapa de teste.*

As Figuras 17 e 18) representam a tentativa de aplicação da técnica de SMOTE (*Sintetic Minority Oversampling Technique) no Random Forests*. Entretanto, como o teste foi feito com classes arbitrárias, não foi possível afirmar se o resultado proporcionou um modelo capaz de predizer classes em áreas não fornecidas na fase de treino.

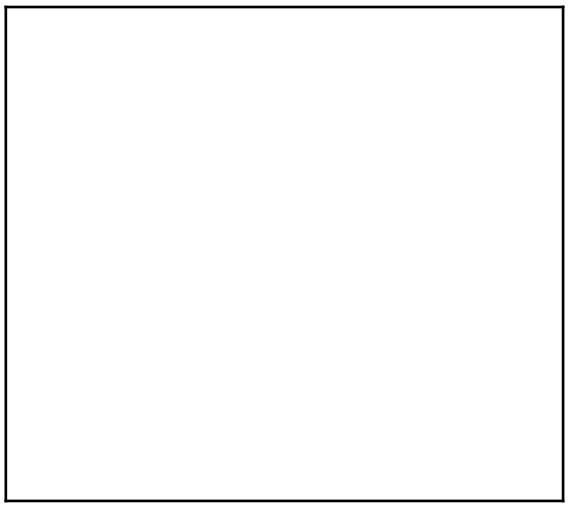

Figura 18 – *Histograma da distribuição das amostras dos dados de treino da classe "Litologia".*

A Figura 19 evidencia, para a variável MDT, uma ótima *feature* de classificação, devida ao fato de haver maior variância dos valores das médias de altitude de cada classe, quando comparado às distribuições amostrais das outras bandas.

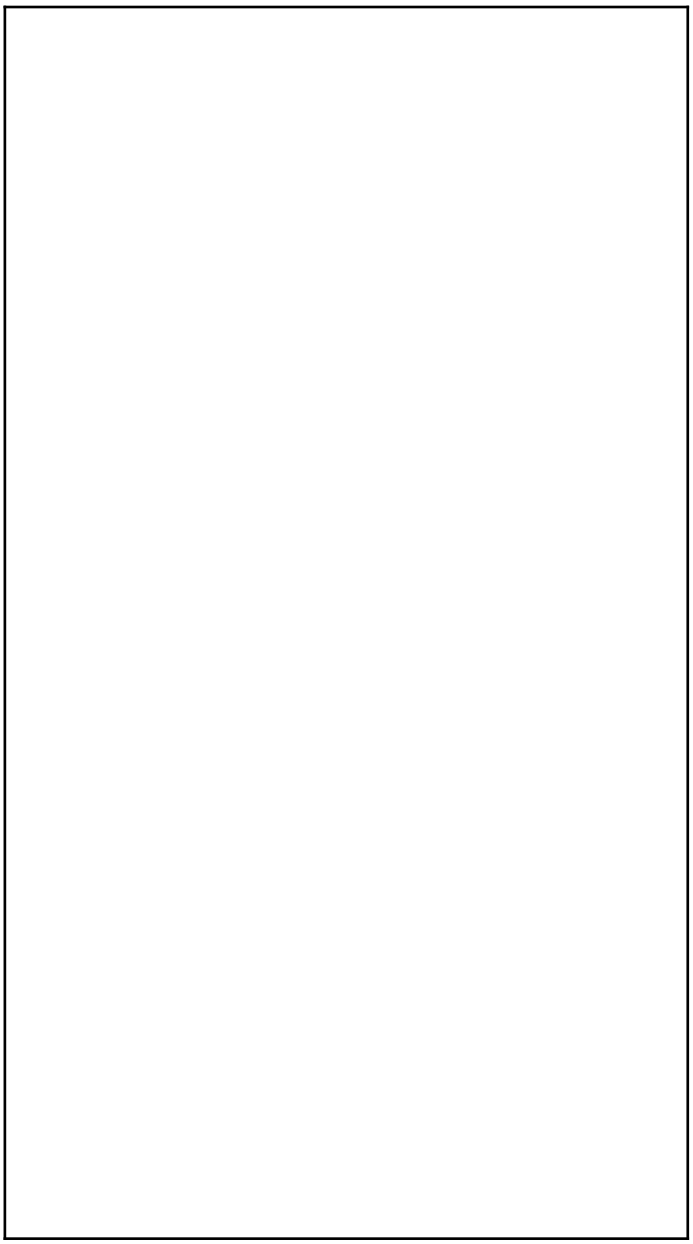

Figura 19 – *Box Plot dos valores do modelo digital de elevação (MDT1), com as médias.*

A Figura 20 deixa evidente a menor distância entre entre as médias dos valores, produzindo uma confusão na etapa de classificação automática. Os valores baixos nos escores dos dados de Urânio equivalente, podem ser explicados pela má atribuição do grau da função polinomial ao se realizar a interpolação pelo método de *splines.*

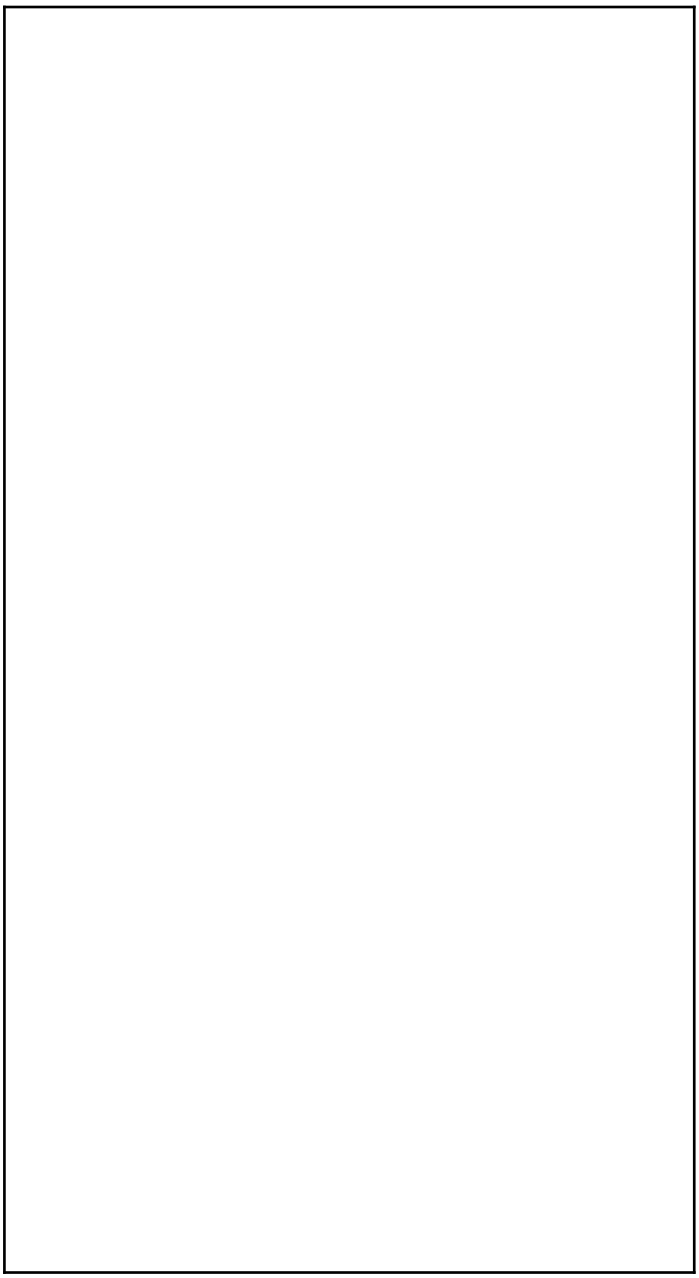

Figura 20 – *Box Plot dos valores amostrados na banda de eU (Urânio equivalente).*

Com as variáveis mostradas na Figura 21 puderam ser identificadas quais são as bandas radiométricas e demais vaariáveis mais importantes para descrever o comportamento das amostras. Os melhores resultados, em ordem de importância, são as *features* de classificação modelo digital de terreno (MDT1), percentual de K (KPERC1) e a razão Tório/Potássio (THKRAZAO).

Figura 21 – *Estes são os escores gerados pela ferramenta Rank do Orange.*

A aplicação do método *Random Forests* em uma subárea do mapa apresentado por [5] (Fig. 22) evidencia as correlações obtidas, mas estes dados não foram ainda correlacionados com as litologias mapeadas, para associação com as rochas grafitosas, o que será desenvolvido na segunda parte da pesquisa.

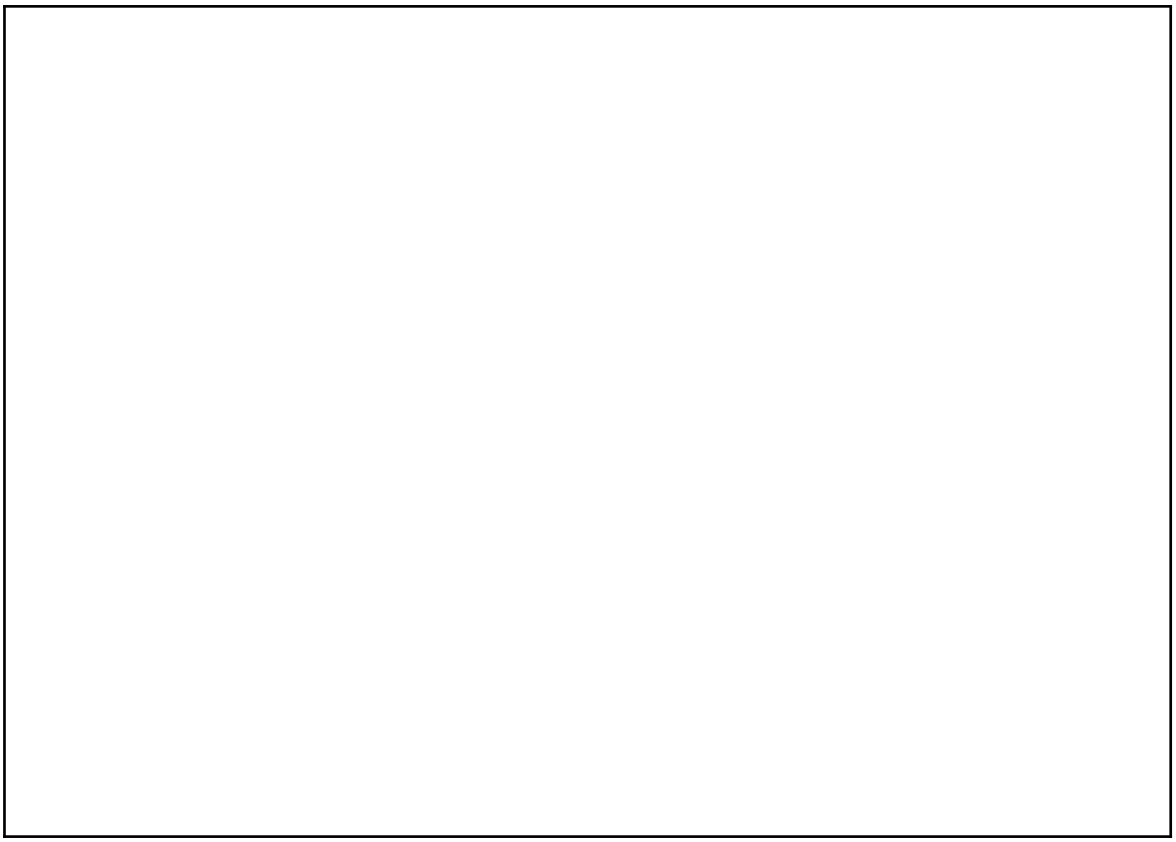

Figura 22 *– Resultado do método Random Forest comparado com o mapa eológico de [5] daa região de Caconde (SP), situada nas proximidades da divisa com o estado de Minas Gerais. Este resuldado ainda não representa bem a capacidade do modelo de predizer resultados de áreas não amostradas pois, para ensinar a máquina foi feita uma replicação excessiva das amostras das classes nas primeira etapa do workflow construído no software Orange.*

O *workflow* do modelo *Random Forests* reproduzindo a técnica SMOTE é apresentado na Figura 23 e o *plot* dos resultados na Figura 24. Com este *plot*, é

possível entender o alto desempenho em reconhecer as classes, pois os dados representam quais foram os pontos inseridos no modelo preditivo como pontos de treinamento. Como houve uma superamostragem da região mapeada, o modelo resultou extremamente preciso para esta área em questão, porém, a predição das classes desta área numa área adjascente não terá o mesmo resultado devido ao viés criado pela super amostragem na área com o mapa geológico na escala de 1:25.000 de [5].

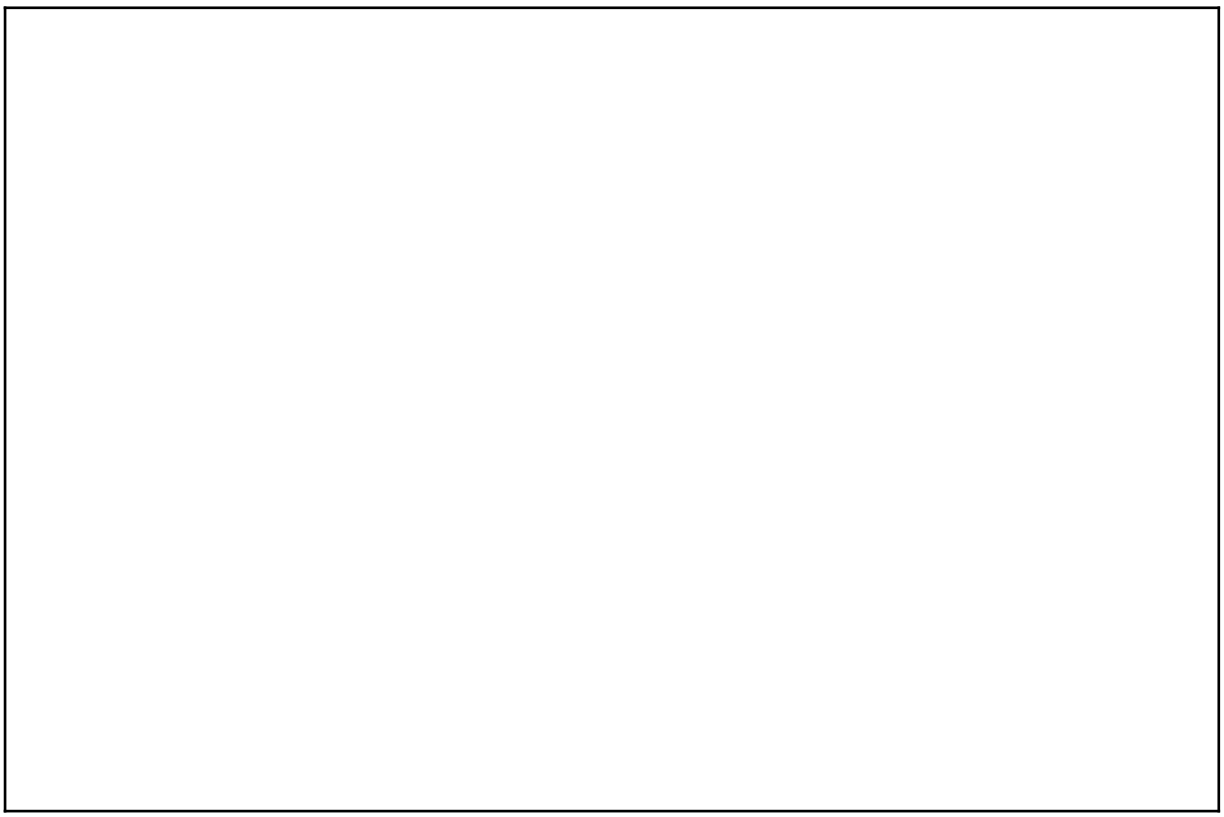

Figura 23 *– Imagem do workflow da tentativa de reprodução da técnica de SMOTE, com os resultados apresentado na Figura 22.*

Figura 24 *– Plot dos pontos de treino do modelo de Random Forests.*

Na Figura 25 pode ser observado que o modelo possui uma alta acurácia nas predições, porém uma baixa precisão. O modelo, ao ser gerado apartir de uma técnica não otimizada de super amostragem/subamostragem, carregou um viéz para a classificação, resultando em classificação muito excessiva as litologias mais amostradas.

A matriz da Figura 26 fornece a informação visual mais clara de como os modelo classificou os pontos. Apesar de ter acertado 100% de muitas classes, houve um grande número de classificações de prováveis falsos positivos provocado pelo viés da superamostragem dos dados.

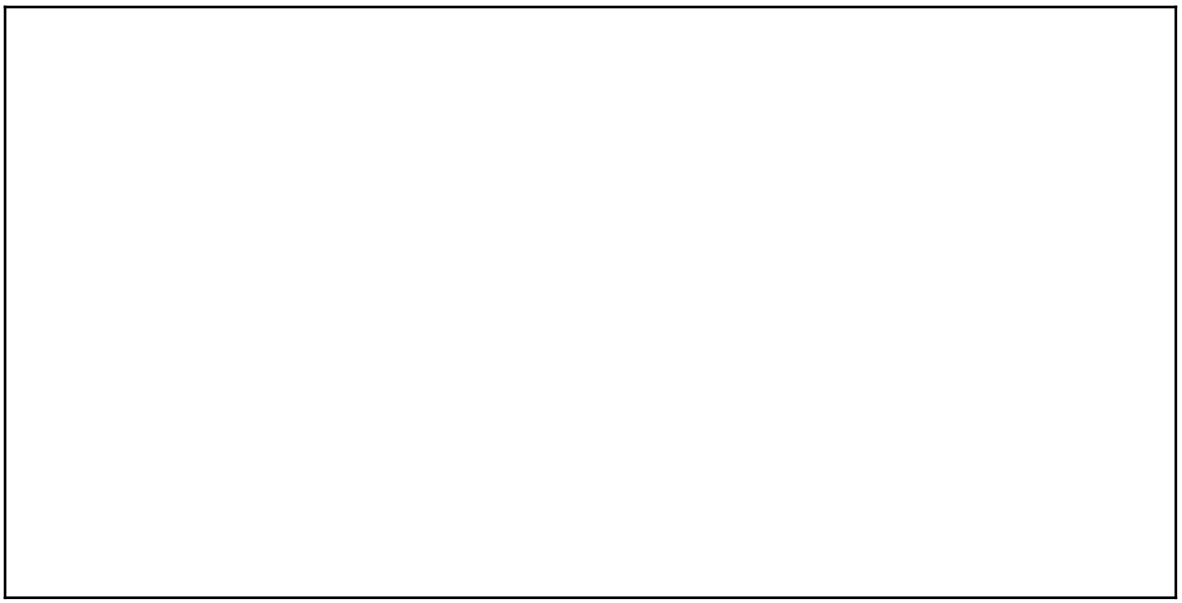

Figura 25 – Matriz de confusão evidenciando a proporção de acerto dos dados preditos.

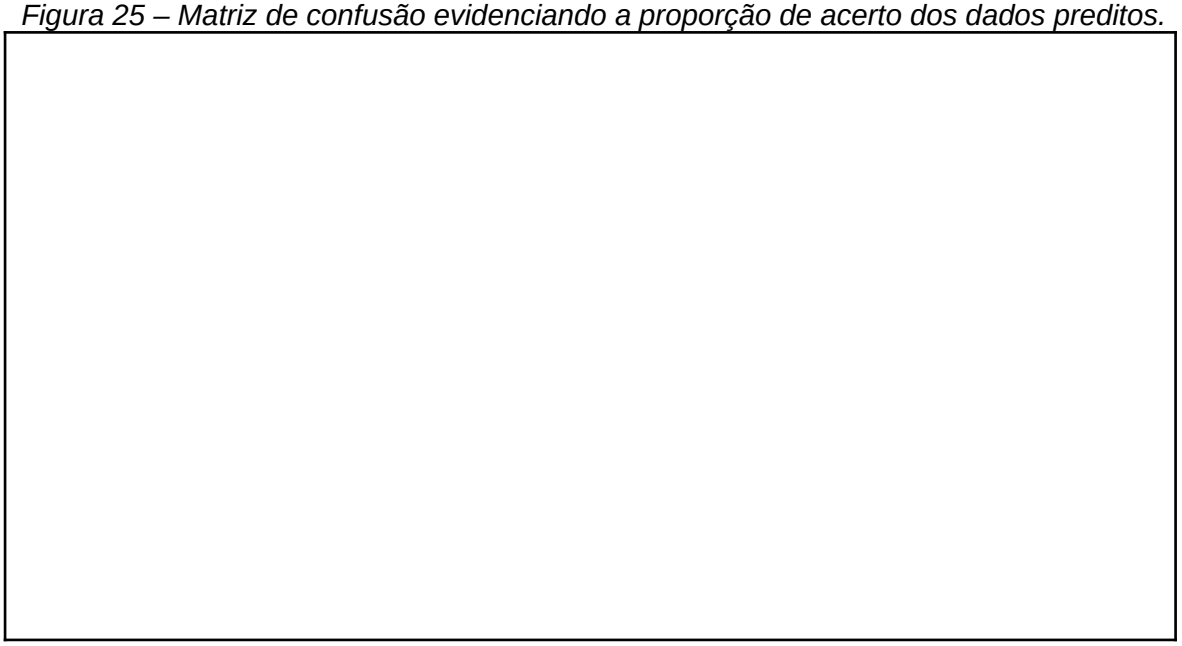

Figura 26 – *Matriz de confusão com relação à proporção dos dados amostrados.*

5.8 Método *Support Vector Machines*

Estudos semelhantes foram feitos para uma outra subárea, na região de Socorro (SP), aplicando-se o *Support Vector Machines*, cujos resultados são apresentados na Figura 27. Os corpos litológicos mapeados por [1] apresentam uma forte orientação e estiramento, devido ao desenvolvimento da zonas de cisalhamento de empurrão de Socorro, desenvolvida em alto grau metamórfico.

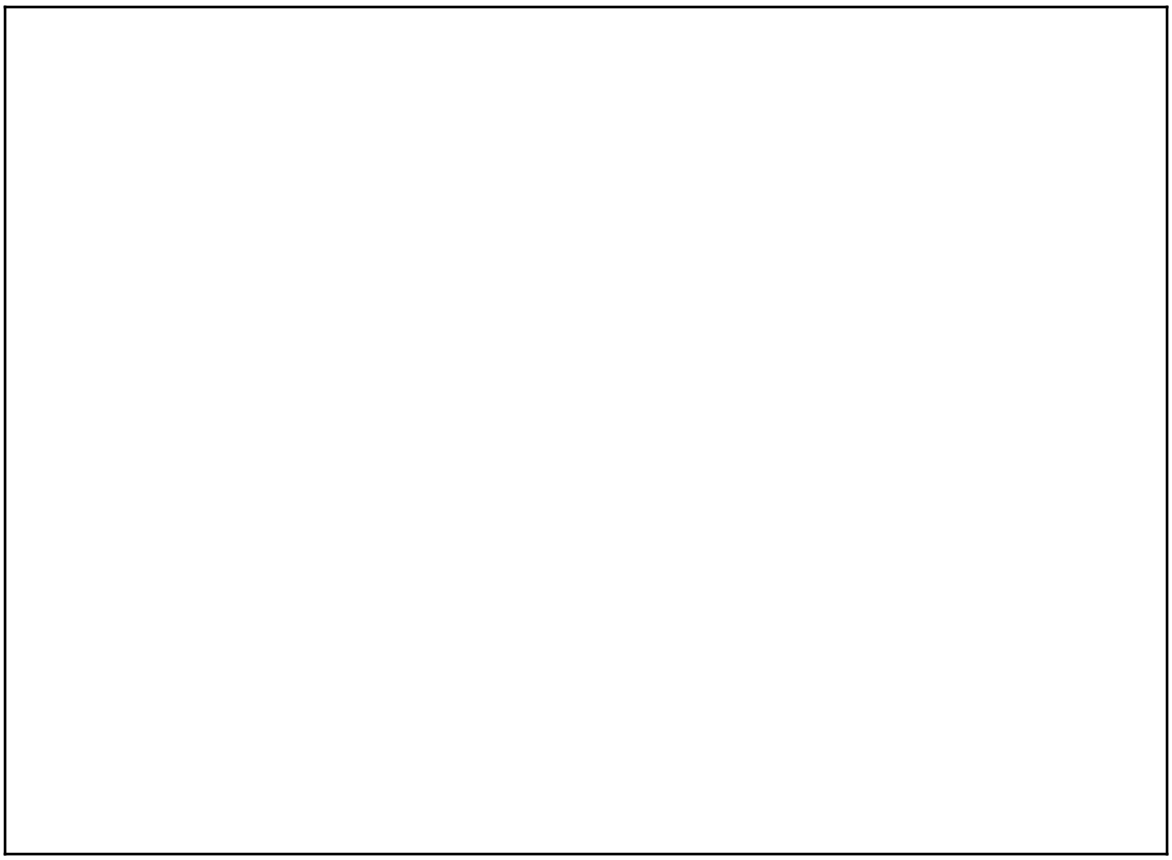

Figura 21 *- Mapa litológico preditivo produzido com o levantamento aerogeofísico "1039" supervisionado pelo pelo mapa geológico de [1] da região de Socorro (SP). A imagem grande à direita representa a tentativa da produção de um mapeamento litológico preditivo não supervisionado.*

Estes resultados serão melhor analisados, mas pode ser notado que, muito embora há correlações evidentes, as extensões dos corpos não são muito correlacionais do do SVM com o mapa geológico e, aparentemente, diversas unidades poderiam ser subdividas.

6 ELABORAÇÃO DE RELATÓRIOS E DE MATERIAIS DIDÁTICOS

No momento em que se dá a entrega deste relatório parcial, já foi iniciada a produção das apostilas no formato *Jupyter Notebook*, formato conhecido pela maneira didática e facilidade na apresentação das linhas de código, e divulgados na plataforma *GitHub no repositório* .

A etapa de pré-processamento de dados já está disponibilizada no repositório *Github* acompanhado do arquivo de texto '*preprocessamento.py*' que contém os

algoritmos de interpolação dos dados discretos precedida pelo tratamento dos arquivos .XYZ de forma automatizada, produzindo um *Xarray* pronto para ser classificado pelos métodos de aprendizagem de máquina no *software Orange* (citar a referência) com os algoritmos de *Support Vector Machine* e *Random Forests.*

7 CONSIDERAÇÕES FINAIS E CONCLUSÕES PRELIMINARES

Apesar dos efeitos da pandemia e de alterações na proposta original decorrentes, que impossibilitaram o estudo de amostras, de um trabalho em campo e de um maior contato entre o orientador e o bolsista, os estudos e, em especial o treinamento nas técnicas de IA alcançaram excelentes resultados.

Com as informações obtidas foi possível entender que as técnicas de classificação automática são promissoras para mapeamento geológioco e, provavelmente, para exploração mineral para grafita, tema este que será melhor abordado na continuidade dos estudos. Porém, é necessário o aprimoramento e refinamento dos processos já feitos bo trabalho, tanto no passo de interpolação quanto nos algoritmos de aprendizagem de máquina, para que se atinja um resultado ótimo e capaz de nos auxiliar no prospecto mineral. Em especial, a correlação ou não com os os dados geológicos de campo deverão ser melhor avaliados, pois estas informações podem por um lado aprimorar os modelos e IA e, por outro lado, contribuir para melhoria dos mapas geológicos e para exploração mineral.

Algo essencial, também identificado como um problema no presente trabalho, é a falta de padronização das informações disponíveis, sendo necessário a criação de um banco de dados padronizado para fácil acesso, e um dos formatos possíveis para a criação deste banco de dados é o formato de *dicionários.*

Outra consideração a se fazer seria a do estudo da linguagem de lógica computacional ‘*Julia’.* Linguagem que recentemente vem ganhando força nos ciências naturais pela eficiencia dos cálculos de grande volume.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] F.C. Freitas, C. Juliani, A. Bustamante, Evolução P-T de rochas metamórficas de alto grau da região de Socorro: Implicações na geração de terrenos

granulíticos de alta pressão no Sudeste do Brasil," *Comun. Geol.*, 99(2): 61–69, 2012.

[2] V.A. Janasi, "Petrogênese de granitos crustais na Nappe de Empurrão Socorro Guaxupé (SP-MG): uma contribuição da geoquímica elemental e isotópica," p. 316, 1999.

[3] V.C.F. Susin, "Interpretação de dados aerogeofísicos para identificação de depósitos de grafita na região de Macarani, no Estado da Bahia," vol. 22, pp. 1– 8, 2019.

[4] L.M.M. Carvalho, "Integração de dados de geofísica aérea aplicada a geologia e à prospecção mineral no Distrito Esmeraldífero de Itabira-Ferros, Quadrilátero Ferrífero, MG," p. 152, 2006.

[5] F.C. Freitas, "Evolução metamórfica dos terrenos granulíticos de Socorro, Caconde (SP) e Cambuí (MG)," *Programa Pós-Graduação em Mineral. e Petrol.*, p. 259, 2006.

[6] K.J. Bergen, P.A. Johnson, M.V. De Hoop, G.C. Beroza, "Machine learning for data-driven discovery in solid Earth geoscience," *Science*, 363(6433), 2019, doi: 10.1126/science.aau0323.

[7] H. Brito, B.S. Piumbini, J.A.M. da Luz, E.M.D. Nascimento, "Caracterização e prospecção de grafita do Complexo Jequitinhonha," *Geol. USP. Série Científica*, 18(1): 67–84, 2018, doi: 10.11606/issn.2316-9095.v18-131162.

[8] E. M. G. Prado, C. R. de Souza Filho, E. J. M. Carranza, and J. G. Motta, "Modeling of Cu-Au prospectivity in the Carajás mineral province (Brazil) through machine learning: Dealing with imbalanced training data," *Ore Geol. Rev.*, vol. 124, p. 103611, 2020, doi: 10.1016/j.oregeorev.2020.103611.

[9] M.J. Cracknell, A.M. Reading, "Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information," *Comput. Geosci.*, 63: 22–33, 2014, doi: 10.1016/j.cageo.2013.10.008.

[10] C. de C. Carneiro, S.J. Fraser, A.P. Crósta, A.M. Silva, C. E. de M. Barros, "Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon," *Geophysics*, 77(4) 2012, doi:

10.1190/geo2011-0302.1.

[11] R. Zuo, E. J. M. Carranza, “Support vector machine: A tool for mapping mineral prospectivity,” *Comput. Geosci.*, 37(12): 1967–1975, 2011, doi: 10.1016/j.cageo.2010.09.014.

[12] A. Melfi, A. Misi, U. Cordani, D. Campos, *Recursos Minerais no Brasil*. 2016.

[13] L. Uieda, (2018). Verde: Processing and gridding spatial data using Green’s functions. Journal of Open Source Software, 3(30): 957, https://doi.org/10.21105/joss.00957