



UNIVERSIDADE FEDERAL DE MINAS GERAIS
(UFMG)
Instituto de Geociências
Departamento de Geologia



TRABALHO GEOLÓGICO DE GRADUAÇÃO

**UTILIZAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA
SUPERVISIONADO PARA MAPEAMENTO GEOLÓGICO: UM
ESTUDO DE CASO NA REGIÃO DE DIAMANTINA, MINAS GERAIS,
BRASIL**

Franco Naghetini

Guilherme Silveira

Orientação: Prof. Dr. Pedro Benedito Casagrande

Coorientação: MSc. Iago Sousa Lima Costa

Belo Horizonte
Novembro/2021

Franco Naghetini

Guilherme Silveira

**UTILIZAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA
SUPERVISIONADO PARA MAPEAMENTO GEOLÓGICO: UM ESTUDO DE
CASO NA REGIÃO DE DIAMANTINA, MINAS GERAIS, BRASIL**

Trabalho Geológico de Graduação apresentado ao
Curso de Geologia da Universidade Federal de
Minas Gerais, como requisito parcial para
obtenção do título Bacharel em Geologia.

Orientação: Prof. Dr. Pedro Benedito Casagrande

Coorientação: MSc. Iago Sousa Lima Costa

Belo Horizonte

Novembro/2021

AGRADECIMENTOS

“Agradeço, primeiramente, aos meus pais, **Maria e Marcos**, por todo amor, carinho, apoio e confiança em todos os momentos da vida. Vocês sempre serão os meus maiores exemplos!

À minha irmã **Camila**, por todo carinho, amor e apoio ao longo da vida e por sempre acreditar em mim. Você é a melhor irmã que alguém poderia querer!

À minha afilhada **Antônia**, por toda docura e amor que emana desde o momento em que chegou ao mundo. O padrinho te ama!

À minha namorada e melhor amiga **K**, por todo amor e carinho e por estar comigo em todos os momentos. Você sempre foi minha maior referência durante a faculdade e fez essa jornada mais leve. Obrigado por fazer cada momento com você tão feliz e inesquecível. Amo você!

Aos meus **avós, tios e primos** por todo amor, carinho e companheirismo. Em especial à minha **Vó Marta**, por ser minha segunda mãe. A senhora sempre estará no meu coração, vó!

À minha terceira mãe **Aletícia**, por todo carinho, paciência e gentileza em todos os momentos.

Aos meus amigos Das Antigas **Vini, Felipe, Fael, Coxa e Amauri** por todo companheirismo e amizade desde sempre. Espero que nossa amizade dure até o restante de nossas vidas!

Aos meus amigos geólogos **Gui, Alice, Léo Bullet, Anna, Bombeiro, Matheus Marlley, Xandão, Tutu, Fê, Má, Primo, Nego Drama, João Lucas, Thaís e Vô** por tornar essa jornada divertida e enriquecedora. Muito obrigado pelas inúmeras horas de estudo, por todos os carnavais e churrascões inesquecíveis e por todos os campos!

Aos meus amigos e colegas da Datamine **Lucas, Gui, Luiz, André, Ju, Maciel, Lúcio e Thum** por sempre me ensinarem, me incentivarem e por sempre acreditarem em mim!

Aos meus professores **Fábio, Aline, Gilberto, Cabral e Alexandre Chaves** que foram essenciais na minha formação como geólogo.

Ao orientador **Prof. Pedro Casagrande** por todo apoio, atenção, amizade e por acreditar na realização deste trabalho. Ao coorientador **Iago Costa** por todas as discussões sobre Aprendizado de Máquina.

Por último, mas não menos importante, à minha dupla **Guilherme** por todos os anos de companheirismo e amizade desde o primeiro dia de faculdade.

Muito obrigado a todos!"

Franco Naghetini

"Eu gostaria de agradecer, inicialmente, aos meus pais, **Ana Lucia e Márcio**, que são, sem dúvida, minha maior fonte de inspiração. Muito obrigado por todo amor, carinho, sacrifícios e cobranças feitas ao longo de toda minha vida. Obrigado pelo incentivo e por acreditarem sempre que a educação é o melhor caminho.

À todos meus familiares, avós, avô, tias, tios, primos e primas por todo carinho e companhia em todos os momentos. Sou o que sou graças a todos vocês!

Aos meus amigos de colégio, especialmente, **Lennon, Bárbara, Evandro, Gabriel e Marcelo**, por todos os momentos divididos nos últimos 10 anos.

Aos meus amigos de faculdade e companheiros de profissão, **Franco, Léo, Anna, Bombeiro, Matheus Marlley, Xandão, Tutu, Fê, Má, Cami, Primo, Nego Drama, Thaís e Vô** por todas as incríveis experiências compartilhadas nesses últimos 6 anos.

À minha namorada e grande companheira de faculdade, **Li**, com quem dividi os melhores e mais bonitos momentos, e de quem guardo as melhores lembranças do período. Obrigado por todo apoio, carinho, paciência.

À todos os funcionários do Instituto de Geociências que fizeram parte da minha formação, especialmente, o Sr. Edison e o Marcelão. Agradeço também a todos os professores por todo conhecimento compartilhado, em especial ao **Prof. Fábio** pela orientação na iniciação científica.

À toda equipe da CERN e da GEOESTRUTURAL onde realizei estágios e tive fantásticas experiências, além de um enorme aprendizado com excelentes profissionais.

Ao orientador **Prof. Pedro Casagrande** pela parceria e apoio desde os primeiros momentos deste trabalho. Ao coorientador **Iago Costa** por todas as contribuições e discussões valiosas.

À Universidade Federal de Minas Gerais por ter aberto todas as portas e ter proporcionado uma vivência única. Ao Instituto de Geociências (IGC) por ter sido minha casa nesses últimos anos.

E finalmente, ao meu grande amigo e parceiro **Franco**, não só por esse trabalho, mas por todos os momentos compartilhados durante a graduação, por todas as trocas de conhecimento, discussões geológicas e por todas as quintas-feiras.

Muito Obrigado a todos!"

Guilherme Silveira

“Essentially all models are wrong, but some are useful.”

- **George Box**

RESUMO

À medida que diversas áreas das Geociências entram na era do *Big Data*, a utilização de técnicas de Aprendizado de Máquina mostra um imenso potencial para a solução de problemas relacionados à identificação de padrões em dados geoespaciais multidimensionais. A aplicação dessas estratégias multivariadas tem apresentado resultados promissores especialmente na confecção de mapas geológicos preditivos a partir de múltiplos dados de sensores remotos. Nesse sentido, o principal objetivo deste trabalho é a aplicação de técnicas de Aprendizado de Máquina Supervisionado para refinamento de mapas geológicos, com a apresentação de um estudo de caso na região de Diamantina, situada no centro-norte de Minas Gerais e no contexto geomorfológico da Serra do Espinhaço Meridional. A partir de uma comparação entre oito algoritmos de Classificação, o modelo de melhor performance foi selecionado, principalmente, com base na métrica *F1-score* e nos mapas de probabilidade por classe e de entropia da informação. As previsões geradas, bem como a influência das variáveis, foram interpretadas a partir da utilização do framework SHAP. Além disso, uma avaliação da Validação Cruzada *K-Fold* (VCKF) como técnica de seleção do modelo de melhor performance foi conduzida, investigando a presença de fenômenos tipicamente associados a dados geoespaciais (*e.g.* correlação espacial). Por fim, foi desenvolvido um repositório remoto aberto que documenta minuciosamente todo o fluxo de trabalho conduzido durante este estudo. Os resultados mostram que os modelos de *Ensemble Learning* (EL) apresentaram as melhores performances, de modo que o modelo *XGBoost* (XGB) gerou previsões com menores incertezas associadas em relação ao *Random Forests* (RF). A interpretação do modelo XGB sugere que as variáveis radiométricas apresentam maiores impactos na previsão das unidades litoestratigráficas. A VCKF mostrou um desempenho adequado na seleção do modelo de melhor performance na presença de correlação espacial, mas na ausência de distorção significativa das distribuições bivariadas entre os conjuntos de treino e teste. O desenvolvimento e a divulgação do repositório remoto representa uma grande contribuição à comunidade geológica, em especial àqueles que buscam ingressar na área de Aprendizado de Máquina aplicado às Geociências.

Palavras-chave: Aprendizado de Máquina Supervisionado, Mapeamento Geológico, Diamantina, SHAP, Validação Cruzada *K-Fold*.

LISTA DE ABREVIATURAS

ACP	Análise de Componentes Principais
ADASYN	<i>Adaptive Synthetic Sampling Technique</i>
AED	Análise Exploratória dos Dados
B02	Banda Landsat 8 - 2
B03	Banda Landsat 8 - 3
B04	Banda Landsat 8 - 4
B06	Banda Landsat 8 - 6
B07	Banda Landsat 8 - 7
CBSO	<i>Cluster-Based Synthetic Oversampling</i>
CT	Contagem Total
DT	<i>Decision Tree</i>
EL	<i>Ensemble Learning</i>
FN	Falso Negativo
FP	Falso Positivo
GB	<i>Gradient Boosting</i>
GPL	<i>General Public License</i>
GT	Gradiente Total
IA	Inteligência Artificial
IBL	<i>Instance-Based Learners</i>
i.i.d.	Independentes e identicamente distribuídos
K	Potássio
KNN	<i>K-Nearest Neighbors</i>
LI	Limiar Inferior
LS	Limiar Superior
MAcgg	Complexo Granito-Gnáissico
MDT	Modelo Digital de Terreno
MG	Minas Gerais
MLP	<i>Multilayer Perceptrons</i>
MMC	<i>Maximal Margin Classifier</i>
MSE	<i>Mean Square Error</i>
NB	<i>Naive Bayes</i>
OLI	<i>Operational Land Imager</i>

OSGeo	<i>Open Source Geospatial Foundation</i>
PC1	Primeira Componente Principal
PP34b	Formação Bandeirinha
PP3csbg	Formação Barão de Guaicuí
PP4egm	Formação Galho do Miguel
PP4esb	Formação Sopa-Brumadinho
PP4esjc	Formação São João da Chapada
QGIS	Quantum GIS
RF	<i>Random Forests</i>
RL	Rregressão Logística
SE	Serra do Espinhaço
SHAP	<i>Shapley Additive Explanations</i>
SIG	Sistema de Informações Geográficas
SLA	<i>Statistical Learning Algorithms</i>
SVC	<i>Support Vector Classifier</i>
SVM	<i>Support Vector Machine</i>
T_a	Conjunto de Treino
T_b	Conjunto de Teste
TBM	<i>Tree-Based Methods</i>
Th	Tório
Th/K	Razão Tório/Potássio
TIRS	<i>Thermal Infrared Sensor</i>
U	Urânio
U/K	Razão Urânio/Potássio
U/Th	Razão Urânio/Tório
UFMG	Universidade Federal de Minas Gerais
VA	Variável Aleatória
VC	Validação Cruzada
VCKF	Validação Cruzada K-Fold
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
VR	Variável Regionalizada
XGB	<i>XGBoost</i>

LISTA DE FIGURAS

Figura 1 - Localização e vias de acesso para a área de estudo.....	3
Figura 2 – O trade-off entre viés (linha contínua azul) e variância (linha contínua laranja). O MSE (linha contínua marrom) e o erro irredutível (linha tracejada preta) são também apresentados. Modificado de JAMES et al. (2013).....	8
Figura 3 – Representação gráfica de uma DT (HASTIE et al., 2009).....	12
Figura 4 - Exemplo de classificação por meio da estratégia MMC. Em um contexto bidimensional, o hiperplano é uma reta. Modificado de JAMES et al. (2013).....	17
Figura 5 - Exemplo de classificação via SVC. A seta de duas pontas indica a largura das soft margins. (A) Alto valor para C . (B) Baixo valor para C . Modificado de JAMES et al. (2013).	19
Figura 6 - Exemplos de SVM com funções kernel não-lineares. (A) SVM com kernel polinomial de grau 3. (B) SVM com kernel de base radial. Modificado de JAMES et al. (2013).	20
Figura 7 - Uma MLP com uma entrada bidimensional, duas camadas ocultas e uma camada de saída. Modificado de BURKOV (2019).....	21
Figura 8 - Exemplo de VCKF com cinco folds, utilizando o MSE como métrica. Figura elaborada pelos autores.....	23
Figura 9 - Contextualização geológica regional da área de estudo destacada pelo retângulo vermelho. Modificado de ALKMIM et al. (2007) e PEDROSA-SOARES et al. (2007).	26
Figura 10 - Coluna estratigráfica das unidades litoestratigráficas nos arredores de Diamantina. As unidades aflorantes na área de estudo encontram-se destacadas em verde. Modificado de LOPES-SILVA & KNAUER (2011).	27
Figura 11 - Mapa integrado modificado 1:25.000 elaborado em 2018 pelos alunos da disciplina Estágio Supervisionado.....	35
Figura 12 - Fluxo de trabalho do projeto.....	39
Figura 13 - Fluxo de trabalho executado para a integração dos sensores remotos.....	40
Figura 14 - Pipeline de pré-processamento dos dados adotado no trabalho.....	44
Figura 15 - Exemplo de ACP, em que os vetores representam as componentes principais (BURKOV, 2019).....	45
Figura 16 – Exemplo de matriz de confusão 6×6 da classe $c2$. VP = verdadeiro positivo, VN = verdadeiro negativo, FN = falso negativo, FP = falso positivo. Figura elaborada pelos autores.....	49

Figura 17 - Superamostragem durante a VCKF em um problema de classificação binária. Apenas o processo na primeira iteração ($k=1$) é apresentado. Figura elaborada pelos autores e baseada em SANTOS et al. (2018).....	50
Figura 18 - Exemplo de um vetor de probabilidades predito para uma determinada instância. Figura elaborada pelos autores.....	52
Figura 19 - Exemplo de summary plot que apresenta o impacto geral das features na predição das classes. Figura elaborada pelos autores.....	54
Figura 20 - Exemplo de summary plot para uma determinada classe. Figura elaborada pelos autores.....	54
Figura 21 - Exemplos de variogramas na ausência (esquerda) e presença (direita) de correlação espacial. Figura elaborada pelos autores.....	55
Figura 22 - Distorção na distribuição bivariada entre os dados utilizados para o treinamento do modelo (cinza) e os dados de teste (rosa). As variáveis geofísicas se encontram estandardizadas. Modificado de HOFFIMANN et al. (2021).	56
Figura 23 - Distribuição das seis unidades litoestratigráficas na área de estudo.....	57
Figura 24 - Distribuição de algumas variáveis independentes por unidades litoestratigráficas.	58
Figura 25 - Matriz de correlação linear entre as variáveis preditoras.	59
Figura 26 - Distribuições das variáveis (A) U (ppm), (B) Th (ppm) e (C) K (%) no Complexo Granito-Gnáissico.....	59
Figura 27 - Distribuições das variáveis (A) K (%) e (B) Th (ppm) na Formação Barão de Guaicuí.	60
Figura 28 - Distribuições das variáveis (A) MDT (m) e (B) U/K na Formação Bandeirinha. 60	
Figura 29 - Distribuições das variáveis (A) U/Th e (B) Th/K na Formação São João da Chapada.	61
Figura 30 - Distribuição da variável U (ppm) na Formação Sopa-Brumadinho.	61
Figura 31 - Distribuições das variáveis (A) U (ppm), (B) MDT (m), (C) K (%) e (D) Th (ppm) na Formação Galho do Miguel.	62
Figura 32 – Distribuição das unidades litoestratigráficas no conjunto de treino após a separação dos dados.....	63
Figura 33 - Distribuição espacial das amostras de treino. As cores representam as unidades litoestratigráficas.	63
Figura 34 - Matriz de correlação linear entre as bandas Landsat 8 estandardizadas e as componentes principais.	65

Figura 35 - Variância explicada relativa de cada uma das cinco componentes principais. A linha preta representa a variância explicada acumulada.....	65
Figura 36 - Matriz de correlação linear entre as variáveis independentes após a redução da dimensionalidade.....	66
Figura 37 - Distribuição das unidades litoestratigráficas no conjunto de treino, após a superamostragem da unidade MACgg.	66
Figura 38 - Scores dos classificadores obtidos na VCKF (superior) e no conjunto de teste (centro). O mapa de calor inferior apresenta as diferenças entre os scores de VCKF e teste. .	68
Figura 39 - Classificadores elencados em ordem decrescente de acordo com os valores de F1-score obtidos na VCKF (superior) e no conjunto de teste (inferior). ..	69
Figura 40 - Matrizes de confusão do conjunto de teste associadas a cada modelo.	70
Figura 41 - Mapas geológicos preditivos e seus respectivos erros de classificação.....	72
Figura 42 - Mapas de probabilidade por classe gerados pelo modelo RF.....	73
Figura 43 - Mapas de probabilidade por classe gerados pelo modelo XGB.	74
Figura 44 - Mapas de entropia dos modelos Random Forests e XGBoost.....	75
Figura 45 - Impacto das variáveis independentes nas previsões do modelo XGBoost.....	76
Figura 46 - Summary plot do Complexo Granito-Gnaissico.....	77
Figura 47 - Summary plot da Formação Barão de Guaicuí.....	77
Figura 48 - Summary plot da Formação Bandeirinha.	78
Figura 49 - Summary plot da Formação São João da Chapada.....	78
Figura 50 - Summary plot da Formação Sopa-Brumadinho.....	79
Figura 51 - Summary plot da Formação Galho do Miguel.....	79
Figura 52 - Distribuições bivariadas dos canais radiométricos agrupadas pelos conjuntos de treino (vermelho) e teste (azul). Essas variáveis apresentam-se estandardizadas.	80
Figura 53 - Variogramas experimentais N-S das variáveis K (vermelho), Th (verde) e U (azul).....	81
Figura 54 - Comparação entre os mapas geológicos integrado (esquerda) e preditivo (direita). ..	84

LISTA DE TABELAS

Tabela 1 - Métricas comumente utilizadas para definição da distância entre instâncias.	11
Tabela 2 - Funções kernel não-lineares mais comuns.	20
Tabela 3 - Informações do Modelo Digital de Terreno utilizado no projeto.....	32
Tabela 4 - Informações dos grids radiométricos utilizados no trabalho.....	32
Tabela 5 - Informações do grid magnetométrico utilizado no trabalho.	33
Tabela 6 - Informações das bandas Landsat 8 utilizadas no projeto.	33
Tabela 7 - Aplicações das bandas Landsat 8 utilizadas.....	34
Tabela 8 - Informações sobre as bibliotecas do Python utilizadas neste projeto.	37
Tabela 9 - Dados resultantes do processo de integração dos sensores remotos.	41
Tabela 10 – Grid de valores avaliados durante a otimização dos hiperparâmetros. Os nomes dos hiperparâmetros seguem a documentação do framework Scikit-Learn.	48
Tabela 11 - Número de instâncias em T_a e T_b após a separação entre treino e teste.	62
Tabela 12 - Sumário estatístico das variáveis independentes estandardizadas no conjunto de treino.....	64
Tabela 13 - Hiperparâmetros ótimos e valores de F1-score para cada algoritmo.	67

SUMÁRIO

1	INTRODUÇÃO	1
1.1	JUSTIFICATIVAS E OBJETIVOS	1
1.2	OBJETIVOS ESPECÍFICOS	2
1.3	LOCALIZAÇÃO DA ÁREA DE ESTUDO	2
2	TEORIA DO APRENDIZADO DE MÁQUINA	4
2.1	APRENDIZADO DE MÁQUINA	4
2.2	APRENDIZADO SUPERVISIONADO E NÃO SUPERVISIONADO	4
2.3	CLASSIFICAÇÃO E REGRESSÃO	6
2.4	VIÉS E VARIÂNCIA	6
2.5	PARÂMETROS DO MODELO E HIPERPARÂMETROS	8
2.6	ALGORITMOS DE CLASSIFICAÇÃO	8
2.6.1	Régressão Logística	9
2.6.2	<i>Naive Bayes</i>.....	10
2.6.3	<i>K-Nearest Neighbors</i>	10
2.6.4	<i>Decision Trees</i>.....	12
2.6.5	<i>Ensemble Learning</i>.....	13
2.6.6	<i>Support Vector Machine</i>	16
2.6.7	<i>Multilayer Perceptrons</i>	20
2.7	VALIDAÇÃO CRUZADA	22
2.8	INTERPRETABILIDADE DE MODELOS	23
2.9	APRENDIZADO DE MÁQUINA NAS GEOCIÊNCIAS	24
3	CONTEXTO GEOLÓGICO	26
3.1	EMBASAMENTO	27
3.2	SUPERGRUPO RIO PARAÚNA	28
3.2.1	Grupo Costa Sena.....	28
3.3	SUPERGRUPO ESPINHAÇO.....	28
3.3.1	Grupo Guinda.....	28
4	BANCO DE DADOS	31
4.1	SENSORIAMENTO REMOTO.....	31
4.1.1	Levantamento Aerogeofísico	31
4.1.2	Landsat 8	33

4.2	MAPA GEOLÓGICO INTEGRADO	34
5	METODOLOGIAS	36
5.1	TECNOLOGIAS UTILIZADAS	36
5.1.1	<i>Softwares de Sistema de Informações Geográficas.....</i>	36
5.1.2	Linguagens de Programação	36
5.1.3	Ambientes de Desenvolvimento.....	37
5.1.4	Sistema de Versionamento.....	38
5.2	FLUXO DE TRABALHO.....	39
5.2.1	Aquisição dos Dados.....	39
5.2.2	Integração dos Sensores Remotos	40
5.2.3	Limpeza dos Dados.....	41
5.2.4	Análise Exploratória dos Dados	42
5.2.5	Separação entre Treino e Teste	42
5.2.6	Pré-Processamento	43
5.2.7	Otimização dos Hiperparâmetros	46
5.2.8	Performance dos Classificadores	48
5.2.9	Mapas Geológicos Preditivos.....	51
5.2.10	Quantificação da Incerteza das Predições	52
5.2.11	Interpretação do Modelo	53
5.2.12	Análise de Fenômenos Associados a Dados Geoespaciais.....	54
6	RESULTADOS	57
6.1	ANÁLISE EXPLORATÓRIA DOS DADOS.....	57
6.2	SEPARAÇÃO ENTRE TREINO E TESTE	62
6.3	PRÉ-PROCESSAMENTO	64
6.4	OTIMIZAÇÃO DOS HIPERPARÂMETROS	67
6.5	PERFORMANCE DOS CLASSIFICADORES.....	67
6.6	MAPAS GEOLÓGICOS PREDITIVOS	71
6.7	QUANTIFICAÇÃO DA INCERTEZA DAS PREDIÇÕES	75
6.8	INTERPRETAÇÃO DO MODELO XGBOOST	75
6.9	ANÁLISE DE FENÔMENOS ASSOCIADOS A DADOS GEOESPACIAIS	80
7	DISCUSSÃO	82
7.1	MAPAS GEOLÓGICOS PREDITIVOS	82
7.1.1	Seleção do Modelo de Melhor Performance.....	82

7.1.2	Análise das Predições das Unidades Litoestratigráficas	83
7.1.3	Desempenho da Validação Cruzada <i>K-Fold</i> na Seleção do Modelo	84
7.2	INTERPRETAÇÃO DO MODELO XGBOOST	85
8	CONCLUSÃO	88
	REFERÊNCIAS BIBLIOGRÁFICAS	89

1 INTRODUÇÃO

1.1 JUSTIFICATIVAS E OBJETIVOS

Este relatório constitui o resultado final do Trabalho Geológico de Graduação, disciplina referente ao décimo período do curso de graduação em Geologia pelo Departamento de Geologia do Instituto de Geociências da Universidade Federal de Minas Gerais (UFMG).

O uso de sensores remotos na Geologia, especialmente associado ao mapeamento geológico, tem se mostrado cada vez mais importante. Dessa forma, com o aumento significativo da quantidade de dados espaciais (*e.g.* imagens de satélite, levantamentos geofísicos), aprimorar as técnicas de utilização e interpretação na busca pelo reconhecimento das litologias tem se tornado fundamental (HARVEY & FOTOPOULOS, 2016).

Em países tropicais e com densas áreas florestadas, como o Brasil, a falta de exposição rochosa, além da dificuldade de acesso se tornam grandes desafios no mapeamento geológico. Desta forma, uma integração adequada de sensores remotos é fundamental para auxiliar na construção de mapas geológicos (COSTA *et al.*, 2019).

A grande quantidade de dados gerados atualmente dificulta a interpretação manual comumente realizada, uma vez que a combinação dos sensores e as relações entre os dados nem sempre são triviais. Usualmente, correlações não-lineares entre variáveis tendem a ser um desafio para a análise multivariada tradicional. Além disso, este tipo de interpretação acaba sendo enviesado pela experiência e conhecimento do intérprete (COSTA *et al.*, 2019; HARVEY & FOTOPOULOS, 2016).

Nesse sentido, a aplicação de técnicas de Aprendizado de Máquina tem se mostrado uma promissora ferramenta na confecção de mapas geológicos preditivos a partir de dados de sensores remotos, como mostrado nos trabalhos de CRACKNELL (2014), CRACKNELL & READING (2014), HARRIS & GRUNSKY (2015), KUHN *et al.* (2019), BACHIRI *et al.* (2019) e COSTA *et al.* (2019).

Nesse contexto, o objetivo principal desse trabalho é a aplicação e discussão de técnicas de Aprendizado de Máquina Supervisionado para mapeamento geológico preditivo, com a apresentação de um estudo de caso em uma área próxima à Diamantina, Minas Gerais, Brasil. Os produtos gráficos gerados podem ser utilizados como um meio de reconciliação entre os

dados e interpretações levantados em uma campanha de mapeamento anterior e os sensores remotos disponíveis. Esses mesmos produtos poderão fornecer *insights* e orientar futuras campanhas que visem o refinamento do mapa geológico da área. Além disso, este trabalho visa estimular a aplicação de técnicas de Aprendizado de Máquina no mapeamento geológico, tendo em vista sua alta relevância e interesse, tanto por parte da academia quanto da indústria.

1.2 OBJETIVOS ESPECÍFICOS

A partir da aplicação de técnicas de Aprendizado de Máquina Supervisionado para mapeamento geológico, este trabalho buscou:

- i. Comparar oito algoritmos distintos e selecionar, dentre eles, o modelo de melhor performance de acordo com as métricas estabelecidas;
- ii. Interpretar as previsões geradas pelo modelo selecionado, a partir da utilização do *framework* SHAP;
- iii. Avaliar o desempenho da Validação Cruzada *K-Fold* como técnica de seleção do modelo de melhor performance, investigando a presença de fenômenos tipicamente associados a dados geoespaciais;
- iv. Desenvolver e disponibilizar de forma gratuita um repositório remoto que documenta minuciosamente todo o fluxo de trabalho desenvolvido com as linguagens abertas Python e Julia.

1.3 LOCALIZAÇÃO DA ÁREA DE ESTUDO

A área deste estudo se localiza na porção centro-norte do estado de Minas Gerais, na mesorregião do Jequitinhonha e microrregião de Diamantina. O polígono¹ que define a área é irregular e abrange três municípios: Datas, Gouveia e Diamantina, de modo que a maior parte da área encontra-se inserida neste último.

A partir de Belo Horizonte, a distância até a área é de aproximadamente 300 km, podendo ser acessada pela BR-040 em direção a Brasília até o encontro com a MG-259, próximo à cidade de Cordisburgo (MG). A partir desse ponto, segue-se pela rodovia estadual no sentido de Diamantina até o acesso a MG-367, que, por sua vez, percorre a área na direção norte-sul (Figura 1).

¹ Coordenadas mínima e máxima, respectivamente: (634163,67; 7969052,06) e (640038,70; 7983240,00).

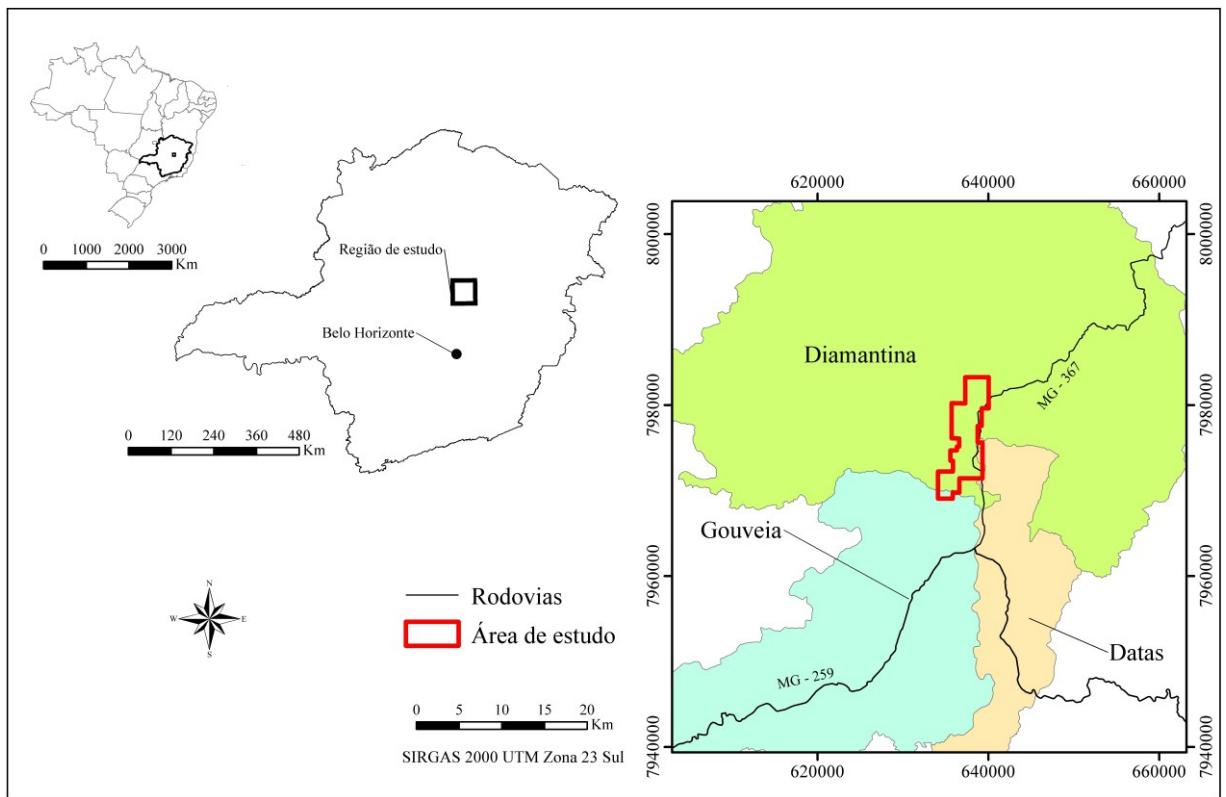


Figura 1 - Localização e vias de acesso para a área de estudo.

2 TEORIA DO APRENDIZADO DE MÁQUINA

2.1 APRENDIZADO DE MÁQUINA

Aprendizado de Máquina (*Machine Learning*) é um termo cunhado por Artur Samuel, um importante pioneiro no campo de jogos e Inteligência Artificial (IA), ao final da década de 1950 (BURKOV, 2019). SAMUEL (1959) postulou a definição para o termo que, segundo ele, se refere à capacidade de um computador aprender sem ser explicitamente programado. Essa capacidade de aprendizado a partir de um conjunto de dados é a chave de todo o Aprendizado de Máquina.

Diversas definições foram desenvolvidas com o objetivo de explicar o processo de aprendizado de máquina, sendo uma das mais conhecidas aquela proposta por MITCHELL (1997):

Dizemos que um agente aprende com a experiência E em relação à tarefa T e a uma medida de performance P , se a performance, medida por P , melhora com a experiência E .

Para EL NAQA & MURPHY (2015), o Aprendizado de Máquina consiste no desenvolvimento de algoritmos capazes de alterar e adaptar sua arquitetura com base em experiências, se tornando melhores para resolver alguma tarefa.

JAMES *et al.* (2013) definem que o Aprendizado de Máquina consiste na estimativa da função f que, quando mapeia um conjunto de variáveis X , fornece informações sistemáticas sobre uma resposta Y , assumindo a existência de uma relação entre X e Y .

$$Y = f(X) + \varepsilon \quad (2.1)$$

em que ε representa um erro aleatório intrínseco.

2.2 APRENDIZADO SUPERVISIONADO E NÃO SUPERVISIONADO

O Aprendizado de Máquina pode ser dividido em diversas categorias, dentre as principais estão o Aprendizado Supervisionado, Não Supervisionado, Semi-Supervisionado e por Reforço (BURKOV, 2019). Contudo, grande parte dos trabalhos publicados relacionados a esse tema reconhecem os dois primeiros tipos como principais (TURNER & CHARNIAK, 2005; JAMES *et al.*, 2013). Segundo BERRY *et al.* (2020), a principal diferença entre os dois

tipos está na existência uma variável dependente² $Y = \{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}$ associada aos dados de treino.

O Aprendizado Supervisionado (*Supervised Learning*), abordagem central do trabalho, é organizado de tal forma que cada instância³ (*i.e.* linha do banco de dados) é representada pelo mesmo conjunto de variáveis independentes⁴ $X = \{X_i, \dots, X_p\}$, sejam elas categóricas, contínuas ou binárias. Se essas instâncias são fornecidas juntamente com rótulos (*labels*) atribuídos por especialistas, sob a forma $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})\}$, então se diz que o Aprendizado é do tipo Supervisionado (KOTSIANTIS, 2007).

Dessa forma, o objetivo do desse tipo de aprendizado é estimar um modelo \hat{f} a partir dos dados de treino $\{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\} \in T_a$, a fim de possibilitar futuras previsões desconhecidas y para novos $x \notin T_a$ (CUNNINGHAM *et al.*, 2018). A estrutura dos dados de entrada (*input*), no Aprendizado Supervisionado, é expressa por uma matriz composta por variáveis independentes X e uma (ou mais) variável independente Y :

X				Y
$x_1^{(1)}$	$x_2^{(1)}$...	$x_p^{(1)}$	$y^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$...	$x_p^{(2)}$	$y^{(2)}$
\vdots	\vdots	\ddots	\vdots	\vdots
$x_1^{(n)}$	$x_2^{(n)}$...	$x_p^{(n)}$	$y^{(n)}$

No Aprendizado Não Supervisionado (*Unsupervised Learning*), por outro lado, para cada i -ésima observação, tem-se apenas um vetor de *features* $x^{(i)}$, mas não uma resposta $y^{(i)}$ associada (JAMES *et al.*, 2013). O objetivo deste tipo de Aprendizado é descrever associações e padrões a partir de um conjunto de instâncias $\{x^{(1)}, \dots, x^{(n)}\}$ (HASTIE *et al.*, 2009). Uma das principais tarefas do Aprendizado Não Supervisionado é a Análise de Agrupamento (*Cluster Analysis*). O principal objetivo desse tipo de análise é verificar a qual grupo (*cluster*) uma determinada instância pertence, dada sua posição no *feature space*.

² Também chamada de variável resposta ou *target*.

³ Também chamada de exemplo, amostra ou observação.

⁴ Também chamadas de variáveis explicativas, variáveis preditoras ou *features*.

(JAMES *et al.*, 2013). A estrutura do *input* nesse tipo de situação é representada como uma matriz de variáveis independentes X :

$$\begin{array}{cccc} & & & X \\ \hline & x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ & x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ & \vdots & \vdots & \ddots & \vdots \\ & x_1^{(n)} & x_2^{(n)} & \cdots & x_p^{(n)} \end{array}$$

2.3 CLASSIFICAÇÃO E REGRESSÃO

As Variáveis Aleatórias (VA) podem ser subdivididas em: (1) contínuas, quando assumem valores numéricos (*e.g.* densidade, teores metalíferos) e (2) categóricas, quando representam classes (*e.g.* litologia, alteração) (JAMES *et al.*, 2013).

Essa distinção é fundamental para o entendimento da diferença entre problemas de Classificação e Regressão. Segundo THAKUR (2020), um problema é dito de Classificação quando a variável dependente é categórica e, por outro lado, caso o *target* seja contínuo, tem-se um problema de Regressão.

BURKOV (2019) ainda subdivide os problemas de Classificação em binário e multinomial, de acordo com o número de classes presentes na variável dependente. A classificação binária é definida quando há somente dois valores possíveis para a variável resposta (0 ou 1). Por outro lado, caso o *target* possua mais de duas classes (*e.g.* 0, 1, 2 e 3), a classificação é do tipo multinomial.

2.4 VIÉS E VARIÂNCIA

No Aprendizado de Máquina, uma função de perda comumente utilizada, *Mean Square Error* (MSE), é definida como:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{f}(x^{(i)}))^2 = E(y - \hat{f}(x))^2 \quad (2.2)$$

em que $y^{(i)}$ representa o valor real para a i -ésima instância e $\hat{f}(x^{(i)})$ o valor predito pelo modelo \hat{f} para essa mesma instância.

O MSE pode ser reescrito como a soma de três componentes fundamentais: variância, viés e erro irredutível, respectivamente (JAMES *et al.*, 2013):

$$E \left(y - \hat{f}(x) \right)^2 = \text{Variância} + [\text{Viés}]^2 + \text{Var}[\varepsilon] \quad (2.3)$$

em que a variância é definida como:

$$\text{Variância} = E \left[(\hat{f}(x) - E[\hat{f}(x)])^2 \right] \quad (2.4)$$

e o viés pode ser expresso como:

$$\text{Viés} = E[\hat{f}(x)] - y \quad (2.5)$$

A variância (*variance*) de um modelo de Aprendizado de Máquina é uma propriedade que informa o quanto esse modelo se modificaria quando seu conjunto de treino sofre modificações. Nesse sentido, quando pequenas alterações nos dados de treino não alteram de forma significativa o modelo \hat{f} , se diz que \hat{f} apresenta baixa variância. Em contrapartida, caso pequenas alterações no conjunto de treino promovam mudanças consideráveis na função estimada, \hat{f} possui altas variância e flexibilidade (HASTIE *et al.*, 2009).

O termo viés (*bias*), por outro lado, consiste no erro que é introduzido quando um problema real é aproximado por um modelo pouco flexível, ou seja, mais simples (JAMES *et al.*, 2013). Em outras palavras, o viés traz informações sobre a capacidade que \hat{f} possui em realizar previsões. Portanto, se um determinado modelo comete muitos erros ao prever valores ou classes para instâncias de treino, diz-se que \hat{f} apresenta alto viés. Por outro lado, \hat{f} possui baixo viés quando comete poucos erros de previsão no conjunto de treino.

De forma geral, ambas as propriedades são inversamente proporcionais entre si, ou seja, quando há uma redução do viés, há um aumento da variância e vice-versa. Esse comportamento, denominado *trade-off* entre viés e variância (Figura 2).

Um determinado modelo, quando apresenta alto viés, está sujeito a um problema denominado subajuste (*underfitting*). Esse fenômeno normalmente ocorre quando o modelo é muito simples para ajustar os dados ou quando as variáveis independentes disponíveis não são suficientemente informativas. Em contrapartida, quando um modelo possui elevada variância, ele está sujeito ao fenômeno de sobreajuste (*overfitting*). Nesse caso, esse problema pode ser

causado por um modelo excessivamente complexo ou quando o número de variáveis independentes é muito superior ao número de instâncias disponíveis (BURKOV, 2019).

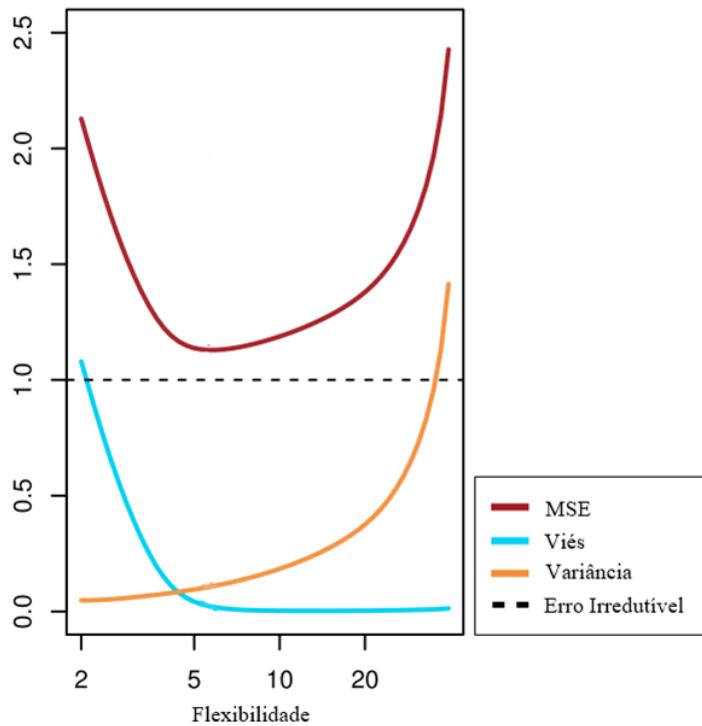


Figura 2 – O trade-off entre viés (linha contínua azul) e variância (linha contínua laranja). O MSE (linha contínua marrom) e o erro irreducível (linha tracejada preta) são também apresentados. Modificado de JAMES et al. (2013).

2.5 PARÂMETROS DO MODELO E HIPERPARÂMETROS

A principal diferença entre essas propriedades está na forma com que elas se relacionam com o modelo. Os parâmetros do modelo participam do seu treinamento e são influenciados pelos dados de treino, de tal sorte que, ao final da etapa de Aprendizado, são definidos a partir da otimização de alguma função objetivo. Os hiperparâmetros, por outro lado, são propriedades do algoritmo, comumente numéricas, que irão influenciar no seu funcionamento. Diferentemente dos parâmetros do modelo, os hiperparâmetros não podem ser otimizados durante a etapa de Aprendizado e, consequentemente, seus valores devem ser previamente definidos pelo analista (BURKOV, 2019).

2.6 ALGORITMOS DE CLASSIFICAÇÃO

Esta seção aborda uma revisão teórica sucinta sobre os algoritmos de Classificação utilizados neste trabalho: Regressão Logística, *Naive Bayes*, *K-Nearest Neighbors*, *Decision Trees*, *Random Forests*, *Gradient Boosting*, *Support Vector Machine* e *Multilayer Perceptrons*.

2.6.1 Regressão Logística

Assim como algumas estratégias de Aprendizado Supervisionado, a Regressão Logística (RL) tem como objetivo estimar o modelo \hat{f} de melhor ajuste que descreva a relação entre as variáveis independentes X e uma variável dependente Y , a partir das instâncias que compõem o conjunto de treino $\{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\} \in T_a$. Diferentemente da Regressão Linear, em que Y é uma variável contínua, na RL, essa variável é binária e segue a Distribuição de Bernoulli (HOSMER & LEMESHOW, 2000; FIGUEIRA, 2006). Essa estratégia, ao invés de modelar o valor de Y diretamente, visa estimar a probabilidade $p(X) = Pr(Y = c | X)$, quando $c \in \{0,1\}$. (JAMES *et al.*, 2013).

Na Regressão Linear, as estimativas para $p(X)$ tendem a não ser razoáveis, já que eventualmente $p(X) < 0$ ou $p(X) > 1$. Nesse sentido, para modelar de forma satisfatória a relação entre $p(X)$ e as variáveis independentes X , é necessária uma função que retorne probabilidades $p(X) \in [0,1]$ para todos os valores assumidos por X . No caso da RL, utiliza-se função logística (JAMES *et al.*, 2013). Para uma determinada instância, em um contexto p -dimensional, essa função é definida como:

$$p(X) = \frac{\exp(\beta_0 + \sum_{j=1}^p \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^p \beta_j x_j)} \quad (2.6)$$

em que β_0 e β_j são os coeficientes (parâmetros) da função.

O método de otimização dos parâmetros, nesse caso, é denominado Máxima Verossimilhança (*Maximum Likelihood*) e visa maximizar a verossimilhança (*likelihood*) do conjunto de treino de acordo com o modelo (BURKOV, 2019). A equação da verossimilhança $L(\beta)$ é representada por:

$$L(\beta) = \prod_{i \in T_a} p(x^{(i)})^{y^{(i)}} \left[1 - p(x^{(i)})^{1-y^{(i)}} \right] \quad (2.7)$$

para valores $y^{(i)} = 0$ ou $y^{(i)} = 1, \forall i \in \{1, \dots, n\}$.

Por conveniência, aplica-se o logaritmo natural em ambos os lados para finalmente obter a função objetivo:

$$\ln [L(\beta)] = \sum_{i \in T_a} (y^{(i)}) \ln[p(x^{(i)})] + (1 - y^{(i)}) \ln[1 - p(x^{(i)})] \quad (2.8)$$

2.6.2 Naive Bayes

As estratégias da família *Statistical Learning Algorithms* (SLA) lançam mão de um modelo probabilístico explícito que informa a probabilidade de cada instância pertencer a uma determinada classe, como é o caso do algoritmo *Naive Bayes* (NB) (KOTSIANTIS, 2007).

O classificador NB, um tipo específico de rede Bayesiana, assume que os atributos preditores são condicionalmente independentes dada uma classe e que todas as variáveis independentes que influenciam na rotulação das classes participam do processo de predição (JOHN & LANGLEY, 1995).

Segundo RISH (2001), considerando uma variável aleatória $C = \{1, \dots, k\}$ que representa os possíveis valores de classe para uma instância e x o *feature vector* que a descreve, o NB associa, a cada uma das classes, uma função discriminante $f_i(x), i \in \{1, \dots, k\}$. Essa função é estimada como a probabilidade à posteriori $Pr(C = i | x)$ que, por sua vez, é obtida pela aplicação do teorema de Bayes:

$$f_i^*(x) = Pr(C = i | x) = \frac{Pr(C = i) Pr(x | C = i)}{Pr(x)} \quad (2.9)$$

Como $Pr(x)$ é igual para todas as classes e não afeta os valores relativos das probabilidades por classe (DOMINGOS & PAZZANI, 1997), esse termo pode ser ignorado e a equação reescrita:

$$f_i^*(x) = Pr(C = i | x) = Pr(C = i) \prod_{j=1}^p Pr(x_j | C = i) \quad (2.10)$$

em que x_j é o valor do j -ésimo elemento do *feature vector* x .

Nesse sentido, o algoritmo NB, por meio do cálculo da Máxima Probabilidade à Posteriori (*Maximum a Posteriori Probability*), classifica uma instância x como pertencente à classe c , se, e somente se, $f_c(x) > f_i(x), \forall i \neq c$ (RISH, 2001; DOMINGOS & PAZZANI, 1997):

$$f_c(x) = \arg \max_{i \in C} Pr(C = i) \prod_{j=1}^p Pr(x_j | C = i) \quad (2.11)$$

2.6.3 K-Nearest Neighbors

Os *Instance-Based Learners* (IBL), também chamados *Lazy Learners*, são algoritmos cuja etapa de aprendizado consiste apenas em armazenar as instâncias do conjunto de treino

(MITCHELL, 1997). Além disso, essa família de algoritmos realiza predições a partir da combinação de instâncias presentes no conjunto de treino e, posteriormente, as descarta (AHA, 1997). Uma das principais vantagens dos IBL é que a função f não é estimada uma única vez para todo o espaço de instâncias, mas sim localmente e de forma distinta para cada exemplo pertencente ao conjunto de teste (MITCHELL, 1997).

O algoritmo não-paramétrico *K-Nearest Neighbors* (KNN), membro da família IBL, assume que instâncias pertencentes a um mesmo conjunto de dados ocorrem próximas umas às outras, desde que apresentem propriedades similares (BURKOV, 2019; COVER & HART, 1967). De forma geral, a classe de uma instância não-rotulada é determinada a partir da classe mais frequente dentre as K instâncias mais próximas a ela no *feature space* (KOTSIANTIS, 2007).

No KNN, a proximidade entre os exemplos é definida por alguma métrica de distância entre amostras $d(x^{(u)}, x^{(v)})$, como a Distância Euclidiana ou a Distância de Manhattan. A Distância de Minkowsky, por sua vez, consiste na generalização dessas duas métricas (KOTSIANTIS, 2007). As equações dessas medidas são apresentadas na Tabela 1, em que $a_j(x)$ é o valor da j -ésima *feature* para a instância x .

Tabela 1 - Métricas comumente utilizadas para definição da distância entre instâncias.

Euclidiana	$d(x^{(u)}, x^{(v)}) = \left(\sum_{j=1}^p a_j(x^{(u)}) - a_j(x^{(v)}) ^2 \right)^{1/2}$
Manhattan	$d(x^{(u)}, x^{(v)}) = \sum_{j=1}^p a_j(x^{(u)}) - a_j(x^{(v)}) $
Minkowsky	$d(x^{(u)}, x^{(v)}) = \left(\sum_{j=1}^p a_j(x^{(u)}) - a_j(x^{(v)}) ^r \right)^{1/r}$

Segundo JAMES *et al.* (2013), o KNN visa estimar a probabilidade condicional de um exemplo ser da classe c , dado seu *feature vector* x_0 , a partir das K instâncias mais próximas a ele representadas pelo subconjunto η_0 :

$$Pr(Y = c | x_0) = \frac{1}{K} \sum_{i \in \eta_0} I(y^{(i)} = c) \quad (2.12)$$

em que $I(y^{(i)} = c)$ é uma variável binária que é igual a 1, se $y^{(i)} = c$ e igual a 0, se $y^{(i)} \neq c$. O exemplo de teste, então, recebe o rótulo da classe com maior valor de probabilidade estimado.

2.6.4 Decision Trees

Os *Tree-Based Methods* (TBM) consistem em uma família de estratégias que visa partitionar o *feature space* em um conjunto de regiões p -dimensionais (HASTIE *et al.*, 2009). Como o conjunto de regras de partição do *feature space* utilizados pelos TBM pode ser representado graficamente como uma árvore invertida⁵, adota-se frequentemente o termo *Decision Tree* (DT) (JAMES *et al.*, 2013).

Uma DT é constituída por nós internos (*internal nodes*), folhas (*leaves*) e um nó raiz (*root node*) que se interconectam por meio de ramos (*branches*). Na Figura 3, os nós internos são representados pelos pontos que partitionam o *feature space* e contém ramos de entrada e saída (*e.g.* $X_2 \leq t_2$), as folhas são as regiões $R = \{R_1, \dots, R_5\}$ e o nó raiz é o único ponto que possui apenas ramos de saída (*i.e.* $X_1 \leq t_1$) (JAMES *et al.*, 2013).

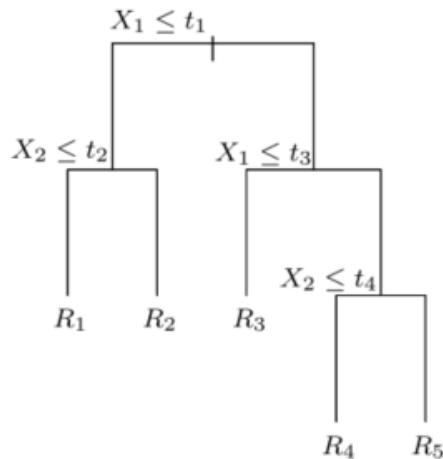


Figura 3 – Representação gráfica de uma DT (HASTIE *et al.*, 2009).

Uma abordagem comumente utilizada para realizar a divisão do *feature space* X em M regiões $R = \{R_1, \dots, R_M\}$ é chamada partição binária recursiva. Ao selecionar uma *feature* X_j e um valor de limiar t , o *feature space* é então subdividido em duas regiões $R_m(j, t) = \{X | X_j < t\}$ e $R_{m+1}(j, t) = \{X | X_j \geq t\}$. Esse processo de partição binária é então repetido

⁵ Estrutura de decisões, cuja raiz situa-se no topo da árvore e as folhas em sua base.

sucessivas vezes até que uma condição de parada seja alcançada (JAMES *et al.*, 2013; HASTIE *et al.*, 2009). Portanto, a árvore é construída recursivamente de cima para baixo, ou seja, se inicia no nó raiz e se desenvolve até as folhas (BREIMAN *et al.*, 1984).

Nos classificadores DT, a seleção de X_j e t para cada nó é realizada com base em medidas de impureza⁶. De forma geral, quanto menor é o valor dessas métricas, maior é a pureza do nó, ou seja, melhor é a capacidade de se separar adequadamente as instâncias pertencentes ao *feature space*. O Índice de Gini G é dado pela equação abaixo, em que \hat{p}_{mc} representa a proporção de instâncias de treino que pertencem à classe c , em que $\hat{p}_{mc} \in R_m$ (JAMES *et al.*, 2013):

$$G = \sum_{c=1}^C \hat{p}_{mc} (1 - \hat{p}_{mc}) \quad (2.13)$$

Por outro lado, a Entropia D é calculada a partir de:

$$D = - \sum_{c=1}^C \hat{p}_{mc} \log_2(\hat{p}_{mc}) \quad (2.14)$$

Após a definição de todas as regiões R , a predição para uma instância de teste \hat{y} pode ser obtida a partir do valor de classe mais frequente dentre todas as instâncias pertencentes a uma determinada região $(x, y) \in R_m$ (JAMES *et al.*, 2013).

2.6.5 Ensemble Learning

O *Ensemble Learning* (EL) é um paradigma de aprendizado, cuja estratégia consiste em treinar diversos modelos de baixa performance e combinar suas predições, visando obter um modelo final de alta performance (BURKOV, 2019). *Bagging* e *Boosting* são estratégias comuns de EL que utilizam DT, ou seja, árvores como *building blocks* para a construção de modelos de alta performance.

Bagging é um método que realiza múltiplas reamostragens com reposição⁷ do conjunto de treino T_α , com o intuito de gerar B subconjuntos de treino e, consequentemente, reduzir a alta variância apresentada pelo algoritmo DT. Em seguida, cada um desses subconjuntos é utilizado para treinar uma DT distinta. A predição $\hat{f}_{bag}(x)$ para uma instância é então

⁶ Os exemplos mais comuns são o Índice de Gini e Entropia.

⁷ Este método é conhecido como *bootstrapping*.

calculada através da média aritmética entre as predições $\hat{f}^b(x)$ de cada modelo (JAMES *et al.*, 2013):

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(x) \quad (2.15)$$

Boosting, por outro lado, é uma abordagem que visa criar iterativamente múltiplos modelos a partir de um modelo simples⁸ (BURKOV, 2019). Assim como no *Bagging*, lida-se com múltiplas árvores $\hat{f}^1, \dots, \hat{f}^B$, porém cada uma delas é gerada utilizando informações de árvores construídas em iterações anteriores (JAMES *et al.*, 2013). Nesse sentido, a principal motivação das estratégias de *Boosting* se baseia na construção de um modelo final de elevada performance a partir da combinação de classificadores fracos (*weak classifiers*) (HASTIE *et al.*, 2009).

2.6.5.1 Random Forests

Random Forests (RF) é um algoritmo derivado do *Bagging* que consiste na combinação de árvores T_1, \dots, T_R , em que cada T_r segue a mesma distribuição e depende das componentes de um vetor aleatório amostrado de forma independente (BREIMAN, 2001).

Inicialmente, um número B de DT é gerado a partir de múltiplas reamostragens com reposição do conjunto de treino. Entretanto, diferentemente do *Bagging*, a cada nó de uma árvore, uma amostra aleatória de m *features* é obtida a partir do conjunto completo de p *features*. Em outras palavras, para cada nó, apenas uma das m features amostradas podem ser selecionadas para particionar o *feature space*, de modo que $m < p$. Como o algoritmo RF limita o número de variáveis independentes candidatas a subdividir o *feature space* a cada partição, as R árvores resultantes tendem a ser fracamente correlacionadas entre si (JAMES *et al.*, 2013). A predição de uma determinada instância é calculada a partir da equação (2.15).

Em geral, o RF está menos sujeito a sofrer um sobreajuste quando comparado ao *Bagging*. Isso se deve ao fato de que o *Bagging*, por considerar todas as variáveis preditoras como candidatas a particionar o *feature space*, gera árvores fortemente correlacionadas entre si, cuja média é incapaz de reduzir consideravelmente a variância do modelo final. Em contrapartida, o RF, por construir árvores descorrelacionadas, tende a apresentar menor variância (BURKOV, 2019; JAMES *et al.*, 2013).

⁸ Algoritmo que é incapaz de reconhecer padrões muito complexos.

2.6.5.2 Gradient Boosting

Gradient Boosting (GB) é um algoritmo de EL que se baseia na ideia de *Boosting*. O modelo f construído pelo GB é treinado de forma aditiva, ou seja, a cada b -ésima iteração, um novo modelo f_b que otimiza uma determinada função objetivo $L(y, f(x))$ é adicionado de modo a melhorar a performance do modelo combinado resultante (CHEN & GUESTRIN, 2016). Segundo BURKOV (2019), pode-se utilizar variações da Máxima Verossimilhança (2.8) como critério de otimização no caso de problemas de classificação.

A princípio, inicializa-se um modelo constante ótimo f_0 que consiste em uma única folha ou nó terminal (HASTIE *et al.*, 2009):

$$f_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y^{(i)}, \gamma) \quad (2.16)$$

em que $y^{(i)}$ representa o valor observado para a i -ésima instância e γ o valor predito. No caso de problemas de classificação, é possível demonstrar que a predição inicial ótima $f_0(x)$ corresponde ao logaritmo natural da chance (*log odds*) de uma classe c :

$$f_0(x) = \ln \left[\frac{Pr(y=c)}{1 - Pr(y=c)} \right] \quad (2.17)$$

Essa predição em *log odds* pode ser facilmente convertida em probabilidade a partir da função logística (2.6).

Após a inicialização do modelo, as seguintes etapas são executadas para cada b -ésima iteração, sendo $b \in \{1, \dots, B\}$ e B o número total de árvores adicionadas ao classificador final (HASTIE *et al.*, 2009):

- i. Cálculo dos pseudo-resíduos ótimos $g_b^{(i)}$ para cada i -ésima instância pertencente ao conjunto de treino:

$$g_b^{(i)} = - \left[\frac{\partial L(y^{(i)}, f_{b-1}(x^{(i)}))}{\partial f_{b-1}(x^{(i)})} \right] \quad (2.18)$$

em que $f_{b-1}(x^{(i)})$ corresponde à predição gerada pelo modelo da iteração anterior para a i -ésima instância.

- ii. Ajuste de uma nova árvore f_b aos pseudo-resíduos $g_b^{(i)}$ e definição das regiões terminais $R_{jb}, j = 1, \dots, J_b$.
- iii. Cálculo dos valores de saída ótimos γ_{jb} para cada uma das regiões terminais $j = 1, \dots, J_b$:

$$\gamma_{jb} = \arg \min_{\gamma} \sum_{x^{(i)} \in R_{jb}} L(y^{(i)}, f_{b-1}(x^{(i)}) + \gamma) \quad (2.19)$$

- iv. Cálculo de novas previsões $f_b(x)$ para as instâncias de treino a partir da atualização do modelo:

$$f_b(x) = f_{b-1}(x) + \sum_{j=1}^{J_b} \gamma_{jb} I(x \in R_{jb}) \quad (2.20)$$

tal que o somatório contempla situações em que uma mesma instância x pertence a mais de uma região R_{jb} .

Ao final das B iterações, as previsões do modelo combinado são dadas por $\hat{f}(x) = f_B(x)$. No caso de problemas de classificação, as etapas descritas acima são realizadas C vezes para cada b -ésima iteração, sendo C o número de classes pertencentes à variável dependente (HASTIE *et al.*, 2009).

Os três principais hiperparâmetros do GB são o número de árvores B , a profundidade das árvores J e a taxa de aprendizado α (BURKOV, 2019). J está relacionado ao número de folhas (*i.e.* regiões terminais) das árvores, enquanto $\alpha \in (0,1)$ pondera a magnitude da contribuição de cada árvore adicionada ao modelo.

O *XGBoost*, um dos classificadores utilizados neste trabalho, é uma das implementações de GB mais poderosas existentes na literatura e foi desenvolvido por CHEN & GUESTRIN (2016).

2.6.6 Support Vector Machine

Support Vector Machine (SVM) é uma estratégia de classificação desenvolvida na década de 1990 que consiste na generalização de um classificador mais simples, denominado *Maximal Margin Classifier* (MMC) (JAMES *et al.*, 2013). Segundo HAYKIN (1999), a abordagem SVM pode ser entendida como uma categoria de redes neurais alimentadas adiante (*feed-forward neural networks*).

O classificador MMC assume que as instâncias pertencentes ao conjunto de treino são linearmente separáveis no *feature space* \mathbb{R}^p por um hiperplano⁹ (BISHOP, 2006). Um hiperplano, em um contexto p -dimensional, é definido por uma combinação linear entre as variáveis independentes $\{X_1, X_2, \dots, X_p\}$ e os parâmetros $\{\beta_1, \beta_2, \dots, \beta_p\}$:

$$\beta_0 + \sum_{j=1}^p \beta_j X_j = 0 \quad (2.21)$$

Nesse sentido, essa fronteira de decisão separa o *feature space* em dois conjuntos, de modo que as observações $x = (x_1, x_2, \dots, x_p)^T$ pertencem a um deles se satisfazem $\beta_0 + \sum \beta_j X_j > 0$ e ao outro se $\beta_0 + \sum \beta_j X_j < 0$. No caso do MMC, o hiperplano escolhido é aquele que maximiza as margens¹⁰, ou seja, a menor distância entre as instâncias de treino e a fronteira de decisão (BISHOP, 2006). As instâncias que definem as margens, ou seja, aquelas mais próximas ao hiperplano são chamadas de vetores de suporte. A Figura 4 ilustra os conceitos de hiperplano, margem e vetores de suporte apresentados acima em um contexto bidimensional.

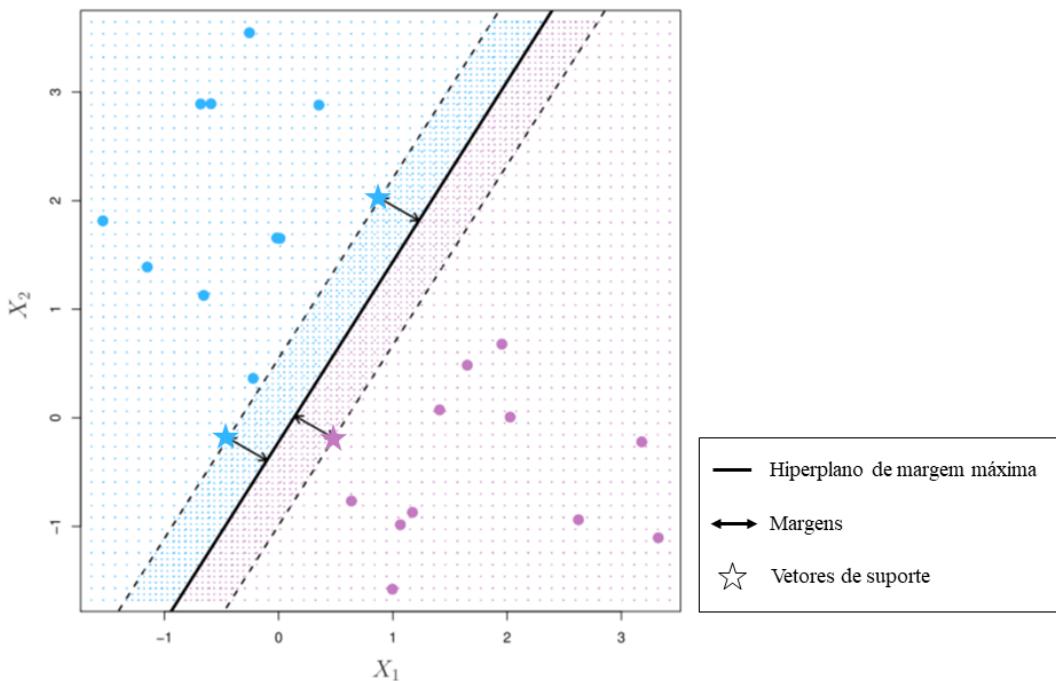


Figura 4 - Exemplo de classificação por meio da estratégia MMC. Em um contexto bidimensional, o hiperplano é uma reta. Modificado de JAMES et al. (2013).

⁹ Formalmente, um hiperplano é um subespaço afim plano de $p - 1$ dimensões.

¹⁰ Esse hiperplano é chamado de hiperplano de margem máxima.

Em muitos problemas reais, as instâncias do conjunto de treino podem não ser perfeitamente separáveis por um hiperplano e, consequentemente, o algoritmo MMC não pode ser aplicado. Nesse sentido, uma generalização do MMC, denominada *Support Vector Classifier* (SVC)¹¹, foi desenvolvida com o intuito de lidar com problemas de classificação de padrões lineares não perfeitamente separáveis (HASTIE *et al.*, 2009).

O SVC, ao invés de calcular o hiperplano que maximiza as margens, permite que algumas instâncias sejam classificadas incorretamente com o intuito de reduzir a variância do modelo e aumentar sua robustez. A construção de um hiperplano que classifica corretamente quase todas as instâncias é possível a partir da utilização das chamadas *soft margins* (JAMES *et al.*, 2013). O SVC pode ser expresso como:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha^{(i)} \langle x, x^{(i)} \rangle \quad (2.22)$$

Em que S representa o subconjunto dos vetores de suporte, $\alpha^{(i)}$ é o parâmetro do i -ésimo vetor de suporte e $\langle x, x^{(i)} \rangle$ é o produto escalar p -dimensional entre uma nova instância e a i -ésima observação de treino.

A largura das *soft margins* e, consequentemente, a tolerância com relação à classificação incorreta das observações é definida pelo hiperparâmetro C (JAMES *et al.*, 2013). Quanto maior é o valor de C , mais tolerante o algoritmo é em relação aos erros de classificação e maior é a largura das *soft margins* (Figura 5A). Por outro lado, valores menores de C resultam em uma menor tolerância e em margens mais delgadas (Figura 5B), o que pode levar o modelo a sofrer um sobreajuste.

¹¹ Também chamada de *Soft Margin Classifier*.

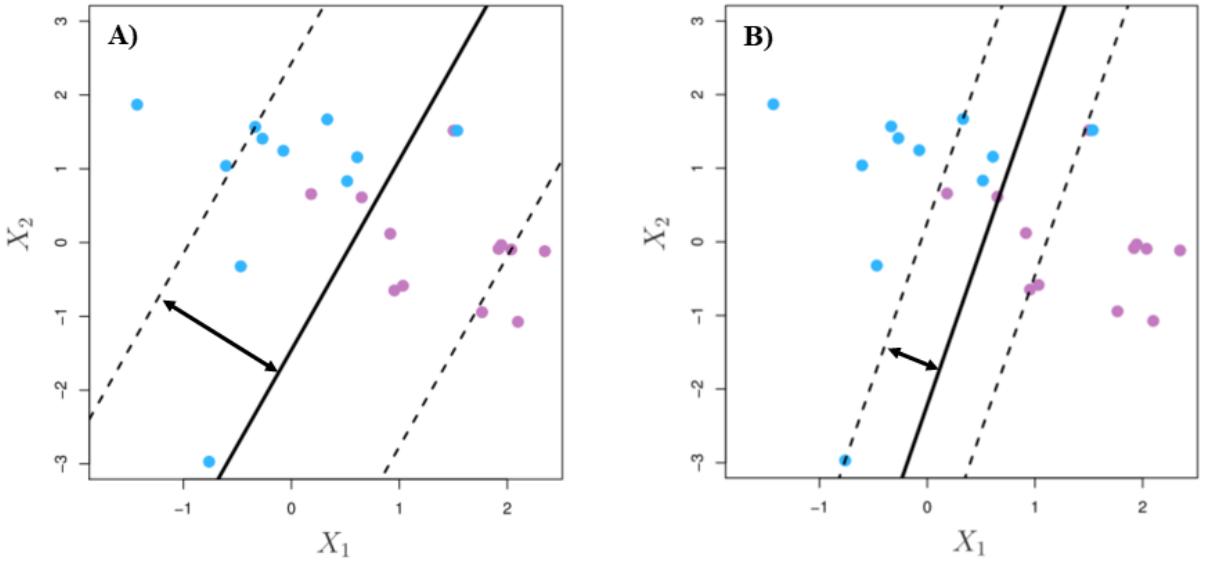


Figura 5 - Exemplo de classificação via SVC. A seta de duas pontas indica a largura das soft margins. (A) Alto valor para C. (B) Baixo valor para C. Modificado de JAMES et al. (2013).

A generalização SVM foi desenvolvida para lidar com problemas cujos padrões não são linearmente separáveis, algo que não é possível a partir da utilização das abordagens MMC e SVC descritas acima. Nesse sentido, a estratégia SVM visa mapear os vetores de entrada x para um *feature space* de elevada dimensionalidade Z a partir de uma função de mapeamento definida à priori. Em seguida, um hiperplano ótimo é construído em Z (VAPNIK, 1995).

De forma geral, como um problema não-linear tem uma maior probabilidade de ser linearmente separável em um espaço de maior dimensionalidade (SMOLA *et al.*, 2000), a ideia principal do classificador SVM é ampliar o *feature space* a partir da utilização de funções *kernel*. A versão não-linear do SVM pode ser definida como:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha^{(i)} K(x, x^{(i)}) \quad (2.23)$$

em que $K(x, x^{(i)})$ é uma função *kernel* entre uma nova instância x e a i -ésima observação de treino $x^{(i)}$. As funções *kernel* mais comuns são a polinomial de grau ω e a função de base radial (*radial basis function*) (JAMES *et al.*, 2013). As equações dessas funções que quantificam a similaridade entre duas observações p -dimensionais x e x' são apresentadas na Tabela 2. Exemplos de SVM com diferentes tipos de *kernel* não-lineares são mostrados na Figura 6.

Tabela 2 - Funções kernel não-lineares mais comuns.

Polinomial	$K(x, x') = \left(1 + \sum_{j=1}^p x_j x'_j\right)^\omega$
Função de Base Radial*	$K(x, x') = \exp\left(-\gamma \sum_{j=1}^p (x_j x'_j)^2\right)$

* γ é uma constante positiva.

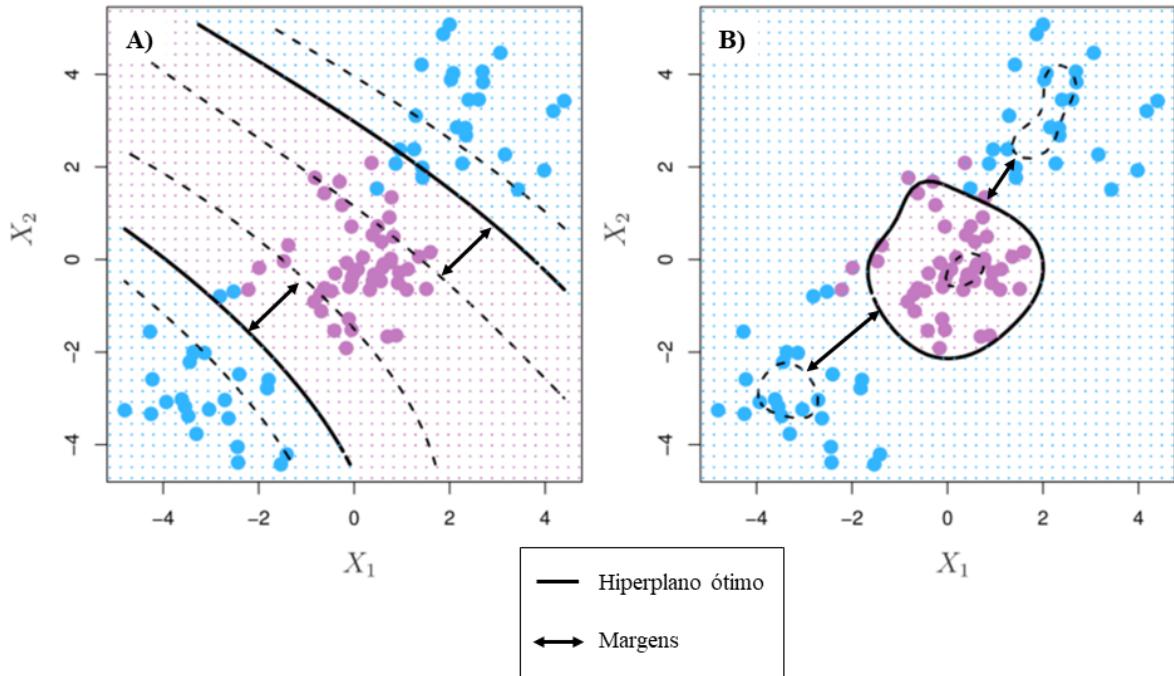


Figura 6 - Exemplos de SVM com funções kernel não-lineares. (A) SVM com kernel polinomial de grau 3. (B) SVM com kernel de base radial. Modificado de JAMES et al. (2013).

2.6.7 Multilayer Perceptrons

Multilayer Perceptrons (MLP) são um tipo específico de redes neurais constituídas por uma camada de entrada (*input layer*), uma ou mais camadas ocultas (*hidden layers*) e por uma camada de saída (*output layer*) (Figura 7). Em outras palavras, essas redes neurais são formadas por um conjunto de unidades (*i.e.* neurônios) interconectadas entre si. Como os sinais se propagam sempre da entrada em direção à saída (e não ao contrário), as MLP também são classificadas como redes neurais alimentadas adiante (HAYKIN, 1999).

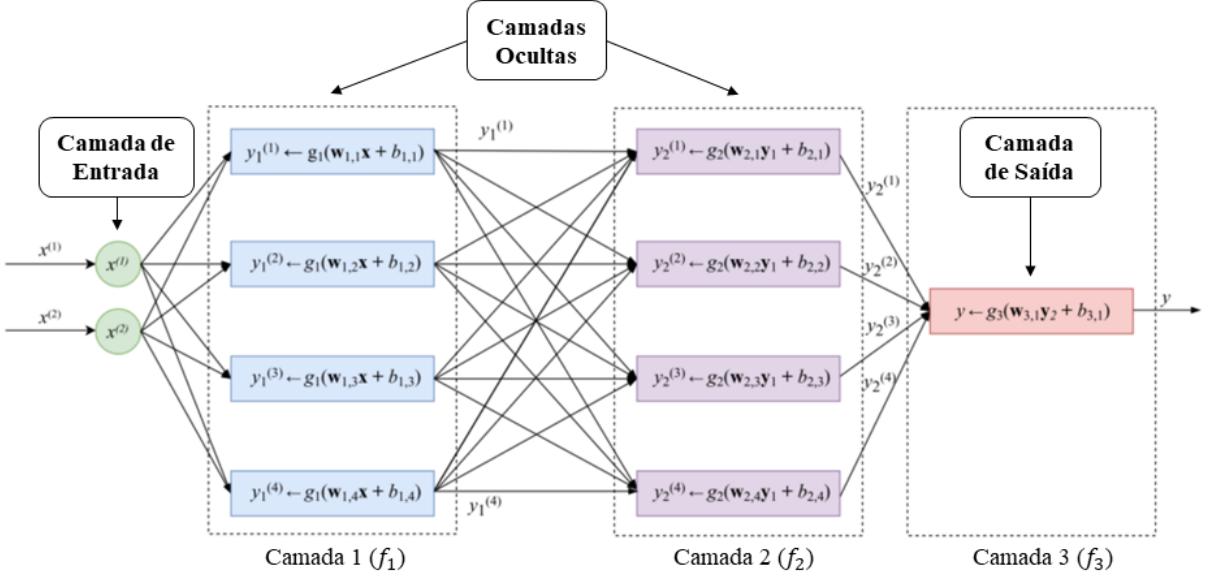


Figura 7 - Uma MLP com uma entrada bidimensional, duas camadas ocultas e uma camada de saída.
Modificado de BURKOV (2019).

Durante a classificação de uma instância, o sinal percorre por toda a rede neural para determinar os valores de ativação de cada unidade de saída. Inicialmente, cada unidade pertencente à camada de entrada recebe um valor de *feature* que é mapeado como um valor de ativação. Esses valores são então combinados na forma de um vetor de entrada e enviados a cada uma das unidades ocultas que fazem conexão. Desse modo, cada unidade oculta aplica, primeiramente, uma transformação linear sobre o vetor de entrada e, em seguida, uma função de ativação g sobre o resultado. Os valores de ativação resultantes são enviados para unidades da camada oculta seguinte ou para neurônios da camada de saída (KOTSIANTIS, 2007).

Segundo BURKOV (2019), uma rede neural (e uma MLP) é uma função matemática aninhada (*nested function*) f_{RN} que apresenta a seguinte forma:

$$f_{RN} = f_L \left(f_{L-1} \left(\dots \left(f_1(x) \right) \right) \right) \quad (2.24)$$

em que L é o número de camadas pertencentes à rede neural. Na equação (2.24), f_L é uma função que retorna um escalar, enquanto f_1, \dots, f_{L-1} são funções vetoriais que podem ser escritas como:

$$f_l = g_l(W_l z + b_l) \quad (2.25)$$

tal que l é o índice de cada camada, g_l é uma função de ativação vetorial não-linear¹², z é um vetor de entrada e W_l (matriz de pesos) e b_l são os parâmetros aprendidos através da otimização de uma função objetivo pelo método gradiente descendente.

Ressalta-se que, assim como a estratégia SVM apresentada anteriormente, as MLP, treinadas via algoritmo *backpropagation*¹³, são capazes de definir uma grande variedade de superfícies de decisão não-lineares (MITCHELL, 1997).

2.7 VALIDAÇÃO CRUZADA

A Validação Cruzada (VC) é uma estratégia de reamostragem comumente utilizada para estimar o erro de generalização¹⁴ de modelos a partir de amostras não vistas por ele (HASTIE *et al.*, 2009). Além disso, como as estimativas do erro de generalização podem ser utilizadas como critério de seleção de modelo, a VC também é adotada na escolha do modelo de melhor performance dentre uma coleção de modelos disponíveis (ARLOT & CELISSE, 2010).

Em muitas situações reais, quando se tem um número escasso de amostras, o conjunto de treino pode ser separado em dois subconjuntos, de modo que um deles é utilizado para treinar o algoritmo e o outro, denominado conjunto de validação, é utilizado para avaliar a performance do modelo (BURKOV, 2019). Nesse sentido, as variedades de VC¹⁵ existentes se diferenciam, essencialmente, na heurística adotada para separar os dados (ARLOT & CELISSE, 2010).

A Validação Cruzada *K-Fold* (VCKF), por exemplo, consiste em uma família de métodos de estimação do erro de generalização em que as instâncias são separadas em K grupos (*i.e. folds*) mutualmente excludentes (HOFFIMANN *et al.*, 2021). Desse modo, a cada k -ésima iteração, o algoritmo é treinado com as instâncias pertencentes às $K - 1$ *folds* e sua performance avaliada na *fold* restante, com base em uma métrica $\mathcal{L}(y, \hat{y})$. A estimativa do erro de generalização (ou da performance) é então calculada pela média aritmética dos resultados de cada iteração (Figura 8):

$$VCKF = \frac{1}{K} \sum_{k=1}^K \mathcal{L}(y, \hat{y})_k \quad (2.26)$$

¹² São exemplos a função logística, tangente hiperbólica (Tanh) e unidade linear retificada (ReLU).

¹³ Um dos algoritmos mais utilizados para se treinar redes neurais alimentadas adiante.

¹⁴ Também chamado de erro de predição.

¹⁵ São exemplos a validação cruzada ordinária (*leave-one-out*) e a validação cruzada k-fold.

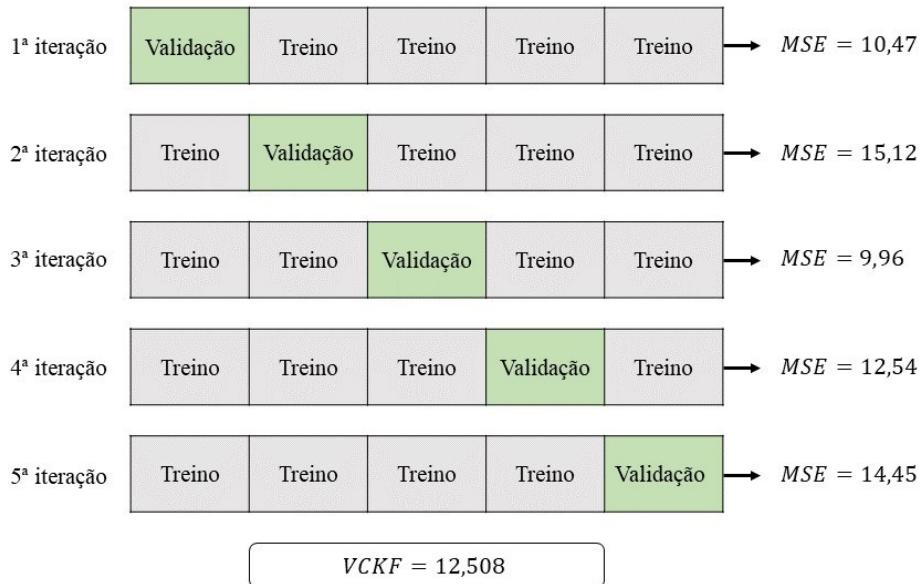


Figura 8 - Exemplo de VCKF com cinco folds, utilizando o MSE como métrica. Figura elaborada pelos autores.

Ressalta-se que as técnicas de VC, essencialmente, assumem duas premissas: (1) as amostras são independentes e identicamente distribuídas (i.i.d.) e (2) as instâncias dos conjuntos de treino e validação são independentes (ARLOT & CELISSE, 2010).

2.8 INTERPRETABILIDADE DE MODELOS

Uma característica importante a ser considerada durante a escolha de um algoritmo de Aprendizado de Máquina é a sua flexibilidade. Em geral, estratégias menos flexíveis fornecem um número mais restrito de formas para a estimativa de f . Por outro lado, algoritmos de maior flexibilidade, como os de *Bagging* e *Boosting*, possibilitam uma grande variedade de formas para a estimativa de f . Embora as estratégias mais flexíveis tendam a fornecer previsões mais acuradas para grandes conjuntos de dados, elas geram modelos mais complexos e de baixa interpretabilidade (JAMES *et al.*, 2013).

Nesse sentido, com o intuito de evitar o *trade-off* entre performance de predição e interpretabilidade, recentemente foram desenvolvidos métodos que auxiliam na interpretação das previsões fornecidas por modelos complexos, tais como LIME (RIBEIRO *et al.*, 2016), DeepLIFT (SHRIKUMAR *et al.*, 2017) e SHAP (LUNDBERG & LEE, 2017).

SHAP (*Shapley Additive Explanations*) é um framework unificado baseado na teoria dos jogos que assinala um valor de importância à cada *feature* para uma predição específica. Os valores de importância, denominados valores de Shapley, foram originalmente propostos por Shapley, na década de 1950, com o intuito de estimar a importância de cada jogador em uma equipe colaborativa (CHEN, 2021).

Segundo LUNDBERG & LEE (2017), uma propriedade importante desse método, a acurácia local, garante que o valor predito pelo modelo de explicação g para uma instância é igual à combinação linear entre os valores de importância (*i.e.* valores de Shapley) atribuídos e os valores simplificados de cada *feature*:

$$g(x') = \phi_0 + \sum_{j=1}^p \phi_j x'_j \quad (2.27)$$

em que $x' \in \{0,1\}_{j=1}^p$ é um vetor de entrada simplificado¹⁶, ϕ_0 é uma constante¹⁷, ϕ_j é o valor de Shapley da j -ésima *feature* e x'_j é o valor simplificado da j -ésima *feature*.

Além disso, o método SHAP considera que cada valor de Shapley indica a magnitude da contribuição de cada variável explicativa na predição de uma instância, seja ela positiva (favorável à predição de uma classe c) ou negativa (desfavorável à predição de uma classe c) (CHEN, 2021). Para uma determinada instância, o valor de Shapley ϕ_j associado à j -ésima *feature* é calculado como:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{j\}}(x_{S \cup \{j\}}) - f_S(x_S)] \quad (2.28)$$

de modo que um modelo é treinado em todos os possíveis subconjuntos de *features* $S \subseteq F$, onde F é o conjunto de todas as *features*. Para calcular a contribuição da j -ésima variável preditora, um modelo $f_{S \cup \{j\}}$ com a j -ésima *feature* inclusa e outro modelo f_S sem ela são treinados. Em seguida, as duas predições são comparadas $f_{S \cup \{j\}}(x_{S \cup \{j\}}) - f_S(x_S)$, onde $x_{S \cup \{j\}}$ representa os valores das *features* de entrada no subconjunto $S \cup \{j\}$ e x_S os valores das *features* de entrada no subconjunto S . Como o efeito de se reter uma variável preditora depende das demais *features* do modelo, a diferença entre as predições é calculada para cada permutação possível de subconjuntos $S \subseteq F \setminus \{j\}$ (LUNDBERG & LEE, 2017).

2.9 APRENDIZADO DE MÁQUINA NAS GEOCIÊNCIAS

Nas últimas décadas, com o aumento de recursos computacionais e o desenvolvimento de tecnologias de sensoriamento remoto modernas, diversas áreas das geociências foram impactadas pela era do *Big Data* (KARPATNE *et al.*, 2018). O crescimento exponencial do número de dados geoespaciais disponíveis favorece a aplicação de técnicas de Aprendizado

¹⁶ x' é a simplificação do vetor de entrada original x calculada a partir de uma função $h_x(x') = x$.

¹⁷ Em problemas de classificação, ϕ_0 é normalmente definido pela probabilidade esperada de uma classe.

de Máquina para a identificação de padrões e integração de dezenas de dados multidisciplinares.

Diversos trabalhos focados na aplicação de técnicas de Aprendizado de Máquina em problemas geocientíficos vêm sendo desenvolvidos, principalmente a partir da última década. Esses estudos abrangem as mais diversas áreas das geociências, tais como: Geotecnia (PURI *et al.*, 2018; CUI & JING, 2019), Recursos Minerais (DUTTA *et al.*, 2010; SAMSON, 2020; CEVIK *et al.*, 2021), Mapeamento Geológico (CRACKNELL, 2014; CRACKNELL & READING, 2014; KUHN *et al.*, 2019; COSTA *et al.*, 2019), entre outros.

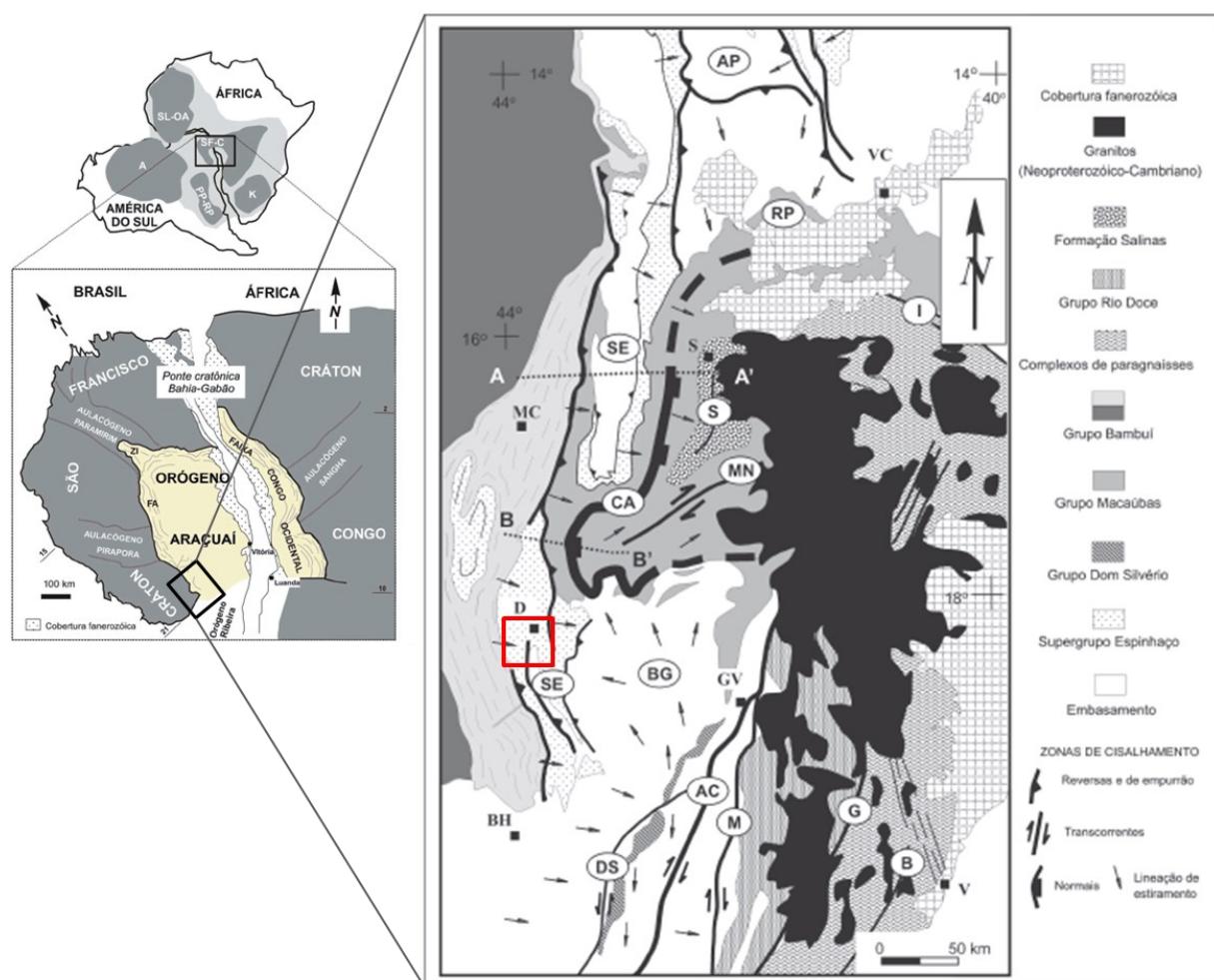
Todavia, a aplicação dessas técnicas para a solução de problemas geocientíficos não pode ser conduzida de maneira direta, uma vez que algumas peculiaridades dos dados geoespaciais violam premissas fundamentais da teoria clássica do Aprendizado de Máquina (HOFFIMANN *et al.*, 2021). Nesse sentido, enquanto a teoria clássica do Aprendizado de Máquina é intimamente ligada às VA, os problemas geoespaciais envolvem um tipo especial de variável denominado Variável Regionalizada (VR). Segundo MATHERON (1975), um fenômeno é regionalizado quando exibe uma certa estrutura espacial¹⁸.

HOFFIMANN *et al.* (2021) afirmam que a utilização de técnicas de reamostragem clássicas (*e.g.* VCKF) em problemas geoespaciais pode gerar estimativas superotimistas do erro de generalização. Na presença de VR, os erros de modelos de Aprendizado de Máquina não podem ser assumidos como i.i.d., em função da existência de correlação espacial e de *trends* causados por processos geofísicos. Esses *trends* promovem distorções (*shifts*) nas distribuições bivariadas entre os conjuntos de treino e teste, principalmente quando esses conjuntos situam-se em áreas geográficas distintas. Nesse sentido, como dados geoespaciais tendem a violar premissas da teoria clássica (*e.g.* amostras i.i.d.), a seleção do modelo de melhor performance via métodos clássicos de reamostragem pode ser inapropriada (FERRACIOLLI *et al.*, 2019).

¹⁸ Também chamada de correlação espacial ou dependência espacial.

3 CONTEXTO GEOLÓGICO

A área de estudo está localizada no sudeste brasileiro e se insere no contexto geomorfológico da Serra do Espinhaço (SE), que se estende por mais de 1.200 km em direção meridiana (KNAUER, 2007). Em termos geológicos, a área se situa no Orógeno Araçuaí, mais especificamente na faixa de dobramentos formada durante o Ciclo Brasiliano (ALMEIDA, 1977), como mostrado na Figura 9.



*Figura 9 - Contextualização geológica regional da área de estudo destacada pelo retângulo vermelho.
Modificado de ALKMIM et al. (2007) e PEDROSA-SOARES et al. (2007).*

A SE pode ser subdividida nos domínios setentrional e meridional. A porção meridional, onde se insere a área de estudo, está situada no estado de Minas Gerais e possui um grande acervo de conhecimento geológico adquirido, principalmente, a partir da descoberta dos depósitos diamantíferos no município de Diamantina (KNAUER, 2007). A área de estudo vem sendo estudada ao longo das últimas décadas pelos alunos e professores que participam da disciplina Estágio Supervisionado. O Centro de Geologia Eschwege, criado ao final da década de 1960, tem desempenhado, por meio da disciplina, um notório papel na

formação de centenas de geólogos no Brasil (FANTINEL, 2005). Dessa forma, a região possui um importante papel para a Geologia do Brasil, em especial, para estudos de mapeamento geológico.

A estratigrafia adotada neste trabalho para a região de Diamantina se baseia nos clássicos trabalhos de PFLUG (1968), SCHÖLL & FOGAÇA (1979) e SCHÖLL (1980). A Figura 10 mostra a coluna estratigráfica das principais unidades.

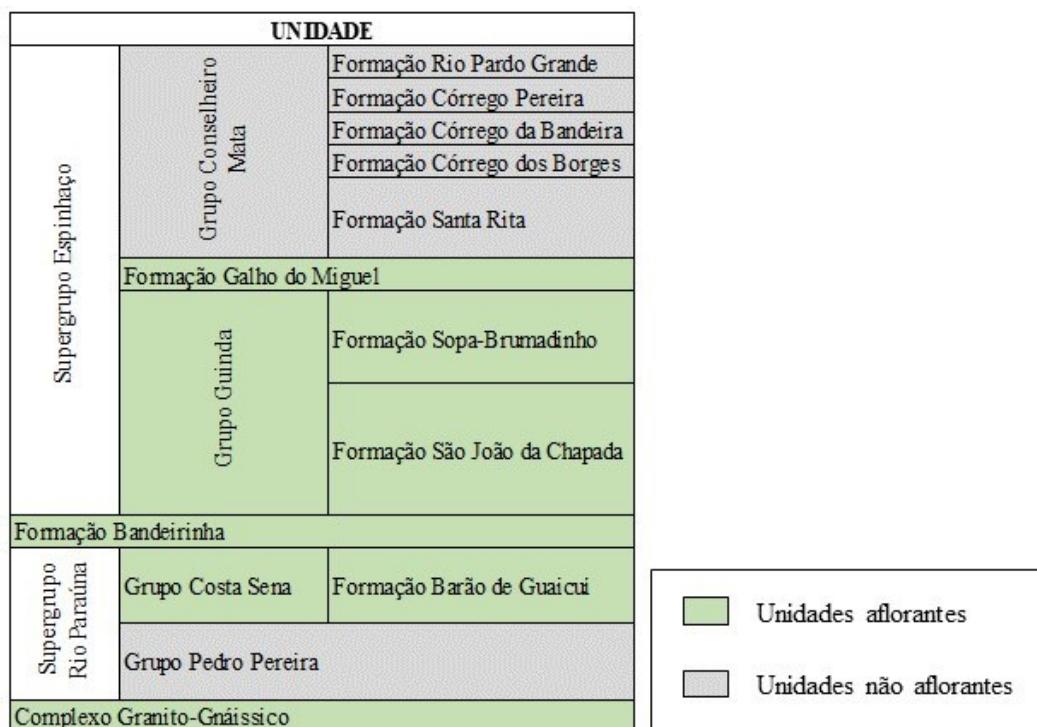


Figura 10 - Coluna estratigráfica das unidades litoestratigráficas nos arredores de Diamantina. As unidades aflorantes na área de estudo encontram-se destacadas em verde. Modificado de LOPES-SILVA & KNAUER (2011).

3.1 EMBASAMENTO

O complexo basal da SE corresponde à sua faixa mediana-central, sendo constituído por rochas graníticas do Complexo Gouveia. Também ocorrem, subordinadamente, rochas gnáissicas (ALMEIDA ABREU, 1995; KNAUER, 1990). Segundo KNAUER (1990), podem ser observadas ocorrências menos expressivas de corpos anfibolíticos e seus produtos milonitizados.

Na região sul da área deste trabalho, afloram rochas de composição granítica a granodiorítica representadas pelo Complexo Granito-Gnáissico (MAcgg). Essas rochas apresentam uma mineralogia composta por quartzo, K-feldspato, plagioclásio, anfibólito, biotita e clorita, além de possuírem textura granular ipidiomórfica a migmatítica. Um

importante aspecto dessa unidade é o elevado grau de alteração das rochas que, muitas das vezes, são identificadas somente como um saprolito arenoso-argiloso rosado e espesso.

3.2 SUPERGRUPO RIO PARAÚNA

Sobreposto ao embasamento, ocorre a unidade denominada Supergrupo Rio Paraúna que, segundo ALMEIDA ABREU (1995), abrange os grupos Pedro Pereira e Costa Sena.

3.2.1 Grupo Costa Sena

3.2.1.1 Formação Barão de Guaicuí

Essa unidade, representada pelo acrônimo PP3csbg, encontra-se nas porções sul e central da área de estudo e aflora comumente como quartzo-sericitic xisto, podendo ainda apresentar biotita e clorita em sua mineralogia essencial e cianita, turmalina, especlarita e magnetita como minerais acessórios. Essa unidade aflora no formato de cristais, em regiões escavadas pela drenagem, e em bordas de morro, com boa exposição. A alteração pode ser descrita como moderada a muito alterada, localmente.

3.2.1.2 Formação Bandeirinha

Situada na porção central da área, a Formação Bandeirinha (PP34b) é composta por quartzitos micáceos, brancos a rosados, pouco alterados. A rocha possui granulação fina, com grãos de quartzo subarredondados a subangulosos e uma matriz fina de cor rósea por conta de presença de óxidos de ferro. As micas brancas definem os planos de foliação. As rochas dessa unidade são típicos exemplos de *red beds*, em que a presença de óxidos de ferro na matriz é responsável pelo tom rosado a avermelhado dos litotipos.

3.3 SUPERGRUPO ESPINHAÇO

3.3.1 Grupo Guinda

3.3.1.1 Formação São João da Chapada

PFLUG (1968) descreve a Formação São João da Chapada (PP4esjc) como quartzitos médios a grossos com estratificações cruzadas, intercalações de seixos na base e lâminas de filito na porção superior. De acordo com SCHÖLL & FOGAÇA (1979), essa formação situa-se na porção basal do Supergrupo Espinhaço, sendo composta por quartzitos, metaconglomerados, filitos e metavulcânicas.

Localmente, a Formação São João da Chapada apresenta-se como uma sequência composta por quartzitos intercalados com filitos sericíticos e hematíticos. A granulação dos quartzitos varia de fina a grossa, podendo conter alguns níveis de quartzito conglomerático. Os filitos hematíticos, por sua vez, são compostos predominantemente por hematita em sua variedade especular. A Formação São João da Chapada é subdividida em três níveis informais A, B e C por SCHÖLL & FOGAÇA (1979).

3.3.1.2 Formação Sopa-Brumadinho

Acima da Formação São João da Chapada, segundo PFLUG (1968), encontra-se a Formação Sopa-Brumadinho (PP4esb) que, por sua vez, é formada principalmente por quartzitos e filitos, com intercalações de metaconglomerados polimíticos e monomíticos. Também foi observado pelo autor lentes de metaconglomerado e intercalações de brecha quartzítica e filito hematítico. A Formação Sopa-Brumadinho é subdividida nos níveis D, E e F por SCHÖLL & FOGAÇA (1979).

Ocorrendo principalmente no limite norte da área, mas não se restringindo a ela, as rochas da Formação Sopa-Brumadinho abrangem filitos sericíticos de coloração verde escura a cinza, com eventuais níveis quartzosos. São observados, ainda, quartzitos bastante heterogêneos de granulação areia fina a grossa, com graus de arredondamento e esfericidade diversos. Nos quartzitos, ocorrem, ainda, estratificações cruzadas tabulares e acanaladas bem preservadas. Destaca-se também a ocorrência de para e orto-metaconglomerados compostos principalmente por seixos de quartzo, quartzito, gnaisses e de conglomerados, além de diamantes.

3.3.1.3 Formação Galho do Miguel

PFLUG (1968) descreve a unidade disposta estratigraficamente acima da Formação Sopa-Brumadinho, a Formação Galho do Miguel (PP4egm) que, por sua vez, é composta por quartzitos puros e finos com estratificações cruzadas e lâminas pouco expressivas de filitos intercaladas. Segundo SCHÖLL & FOGAÇA (1979), essas estruturas são descritas como megaestratificações cruzadas.

Ocorrendo nas bordas nordeste e sudoeste da área, a Formação Galho do Miguel é constituída por ortoquartzitos de granulação areia fina, com níveis locais de sericitita esverdeada. As rochas possuem coloração branca, por vezes róseas a amarela, de modo que megaestratificações cruzadas > 5 m são estruturas diagnósticas dessa unidade.

Ressalta-se que as descrições locais das unidades litoestratigráficas da área de estudo são resultado do mapeamento na escala de 1:25.000 realizado por alunos do curso de Geologia da UFMG, durante a disciplina Estágio Supervisionado ofertada no ano de 2018.

4 BANCO DE DADOS

O banco de dados utilizado para o desenvolvimento deste trabalho é constituído por dados de sensores remotos¹⁹ (*i.e.* gamaespectrométricos, magnetométricos e Landsat 8) e por um mapa geológico²⁰ integrado 1:25.000 da área. Os mapas dos sensores remotos utilizados podem ser encontrados no Anexo I.

4.1 SENSORIAMENTO REMOTO

As definições para Sensoriamento Remoto foram postuladas ao longo dos anos e podem ser resumidas como:

A utilização conjunta de sensores, equipamentos para processamento de dados, equipamentos de transmissão de dados colocados a bordo de aeronaves, espaçonaves, ou outras plataformas, com o objetivo de estudar eventos, fenômenos ou processos que ocorrem na superfície do planeta Terra a partir do registro e da análise das interações entre a radiação eletromagnética e as substâncias que o compõem em suas mais diversas manifestações (NOVO, 2010).

Na geologia, os dados de sensores remotos podem ser aplicados em diversas áreas e escalas, desde estudos ambientais, como o monitoramento de áreas de desmatamento e queimadas, até a Prospecção Mineral (MARTINS E SOUZA FILHO *et al.*, 2006).

4.1.1 Levantamento Aerogeofísico

Os dados geofísicos utilizados nesse trabalho foram adquiridos no início dos anos 2000 pela empresa Megafísica Survey Aerolevantamentos S.A e obtidos através da biblioteca da UFMG. Esse levantamento faz parte do Programa de Levantamento Aerogeofísico de Minas Gerais coordenado pela Secretaria de Estado de Minas e Energia do Governo de Minas Gerais. Ao todo, os levantamentos abrangem 7.000 km em perfis e uma área de 1.567 km² (CODEMIG, 2000).

Realizado na área denominada Faixa São João da Chapada – Datas, o levantamento contou com linhas de voo de direção N20E espaçadas entre si de 250 metros (CODEMIG, 2000). As linhas de voo foram interpoladas com uma resolução de 62,5 metros x 62,5 metros pela própria empresa responsável pelo levantamento.

Nesse trabalho, foram usados *grids* radiométricos (*i.e.* Contagem Total, Urânio, Tório, Potássio e suas razões), magnetométricos (*i.e.* Gradiente Total) e um Modelo Digital de Terreno, cujas informações são apresentadas pela Tabela 3.

¹⁹ <https://github.com/fnaghetini/Mapa-Preditivo/tree/main/data/raster>

²⁰ <https://github.com/fnaghetini/Mapa-Preditivo/blob/main/shp/lithology.shp>

Tabela 3 - Informações do Modelo Digital de Terreno utilizado no projeto.

Atributo	Sigla	Unidade	Sistema de Referência	Resolução Inicial
Modelo Digital de Terreno	MDT	m	Córrego Alegre 1970-72 UTM Zona 23S	62,5 m

O levantamento aerogamaespectrométrico é uma técnica de levantamento de dados radiométricos focada na identificação da radiação gama emitida pelas rochas que compõem a superfície, mais especificamente da radiação gerada pelo decaimento dos isótopos ^{40}K , ^{238}U e ^{282}Th hospedados em rochas superficiais (TELFORD *et al.*, 1990).

De acordo com MINTY (1988), a radiação gama no solo é inversamente proporcional à densidade do meio que atravessa, sendo assim, a radiação captada pelo sensor corresponde a, no máximo, 30 a 40 cm de profundidade. Logo, os dados coletados correspondem à cobertura mais superficial da área estudada.

Ao todo, foram utilizados sete *grids* radiométricos do levantamento Faixa São João da Chapada – Datas (Tabela 4).

Tabela 4 - Informações dos grids radiométricos utilizados no trabalho.

Atributo	Sigla	Unidade	Sistema de Referência	Resolução Inicial (m)
Potássio	K	%		
Tório	Th	ppm		
Urânio	U	ppm		
Contagem Total	CT	$\mu\text{R/h}$	Córrego Alegre 1970-72 UTM Zona 23S	62,5 m
Razão Urânio/Potássio	U/K	-		
Razão Tório/Potássio	Th/K	-		
Razão Urânio/Tório	U/Th	-		

A Magnetometria é um importante método potencial utilizado na busca de depósitos minerais metálicos, sendo a prospecção de depósitos de ferro sua principal aplicação. No mapeamento geológico, esse tipo de levantamento pode auxiliar significativamente na identificação das rochas que contenham minerais magnéticos, além de ser uma importante ferramenta no mapeamento de estruturas em profundidade (KEAREY *et al.*, 2009).

Dentre os dados adquiridos e processados pela empresa Megafísica Survey Aerolevantamentos S.A., optou-se somente pelo uso do *grid* de Gradiente Total (Tabela 5).

Tabela 5 - Informações do grid magnetométrico utilizado no trabalho.

Atributo	Sigla	Unidade	Sistema de Referência	Resolução Inicial
Gradiente Total	GT	nT/m	Córrego Alegre 1970-72 UTM Zona 23S	62,5 m

4.1.2 Landsat 8

Sendo o mais novo satélite da família até o momento da realização deste trabalho, o Landsat 8 foi lançado em 11 de fevereiro de 2013, é formado, essencialmente, por dois instrumentos, *Thermal Infrared Sensor* (TIRS) e *Operational Land Imager* (OLI) e fornece uma cobertura de todo o planeta, com uma resolução temporal de 16 dias (PARAMOS FILHO *et al.*, 2021). O Landsat 8 possui uma resolução espacial de 30 metros nas bandas visíveis NIR e SWIR (NASA, 2021).

Possuindo 11 bandas multiespectrais e uma banda pancromática, o Landsat 8 é o satélite com mais bandas dentre todos da família, das quais nove estão relacionadas ao sensor OLI e duas ao TIRS.

Neste trabalho, foram utilizadas cinco bandas multiespectrais do Landsat 8, cujas informações são apresentadas na Tabela 6. As aplicações desses dados são summarizadas na Tabela 7.

Tabela 6 - Informações das bandas Landsat 8 utilizadas no projeto.

Atributo	Sigla	Unidade	Sistema de Referência	Resolução Inicial
Banda 2	B02	NC*		
Banda 3	B03	NC*		
Banda 4	B04	NC*	WGS84	30 m
Banda 6	B06	NC*		
Banda 7	B07	NC*		

*Níveis de cinza. As bandas Landsat 8 disponibilizadas podem apresentar até 55.000 níveis de cinza (USGS, 2021).

Tabela 7 - Aplicações das bandas Landsat 8 utilizadas²¹.

Banda	Aplicações
B02/B03/B04	As bandas 4, 3 e 2 correspondem aos sensores na faixa do espectro visível, mais conhecidas como bandas R (<i>red</i>), G (<i>green</i>) e B (<i>blue</i>), respectivamente. Quando combinadas, podem ser usadas em diversos estudos de uso e ocupação.
B06/B07	As bandas 6 e 7 cobrem diferentes fatias do infravermelho de ondas curtas e são utilizadas na identificação de terras úmidas e secas, além da aplicação em geologia na identificação de rochas e solos.

Dentre as cinco bandas utilizadas, destacam-se aquelas relacionadas à faixa infravermelho próximo (*i.e.* bandas 6 e 7), por captarem a faixa de comprimento de onda relacionada aos solos e as rochas. Essas bandas encontram-se disponíveis online de forma gratuita²².

4.2 MAPA GEOLÓGICO INTEGRADO

O mapa utilizado como referência para a realização deste trabalho é resultado da disciplina de Estágio Supervisionado²³, oferecida no oitavo período do curso de Geologia da UFMG. Essa disciplina é parte do currículo obrigatório dos estudantes e é realizada anualmente no Centro de Geologia Eschwege, na cidade de Diamantina (MG).

O mapa integrado modificado na escala de 1:25.000 utilizado neste trabalho (Figura 11) foi confeccionado no ano de 2018. Ao todo participaram 24 discentes, divididos em seis grupos de mapeamento:

- i. Grupo 1 – Flávia Maria Faria de Souza Lima, Jéssica Barbosa Amorim, Fernando Pletschette Galvão e Breno Campo Magalhães;
- ii. Grupo 2 – Camila Miranda Brighenti, João Lucas Andrade Penna, Matheus Luís de Sales Oliveira e Thaís Keuffer Mendonça;
- iii. Grupo 3 – Alberto Vital Dias Duarte, Carlos José Lopes Rodrigues, Giulia Marina Cerqueira Dias e Henrique Borgatti;
- iv. Grupo 4 – Carlos Eduardo Vieira, Gabriel Costa Alvarez, Matheus Alonso Castelo Pena e Victor Costa de Souza e Silva;

²¹ Modificado de: <https://landsat.gsfc.nasa.gov/landsat-8/landsat-8-bands>.

²² <https://earthexplorer.usgs.gov>

²³ Código GEL663.

- v. Grupo 5 – Bruno Pandolf Ladeira, Calistrato Lopes de Muros, Mayá Quaresma e Silva e Vittor Clementino de Souza Javidan;
- vi. Grupo 6 – Clara Moreira Moraes, Felipe Augusto Alves Pereira, Gilberto Mendes de Cunha Júnior e Lucas Lana de Paula.

Os grupos foram orientados pelos professores Dra. Aline Tavares de Melo, Dr. Gabriel Jubé Uhlein e Dr. Jorge Geraldo Roncato Júnior.

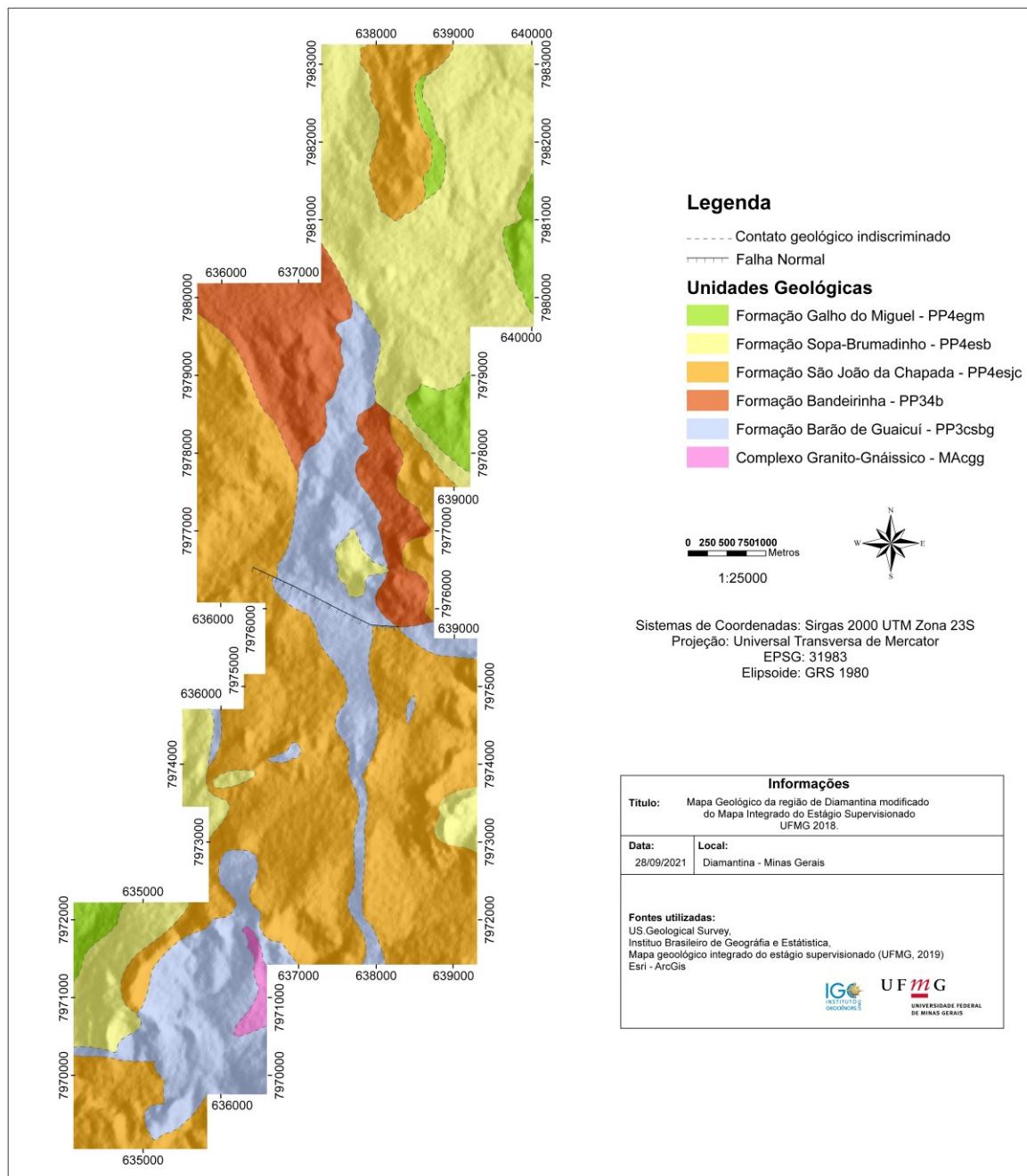


Figura 11 - Mapa integrado modificado 1:25.000 elaborado em 2018 pelos alunos da disciplina Estágio Supervisionado.

5 METODOLOGIAS

5.1 TECNOLOGIAS UTILIZADAS

A crescente utilização de diferentes tecnologias no campo das Geociências, especialmente na Geologia, está relacionada à necessidade de suprir as principais demandas da área (DE SOUZA FILHO & CRÓSTA, 2003).

Além dos amplamente aplicados *softwares* de Sistema de Informações Geográficas (SIG), este trabalho dispôs da utilização de diferentes tecnologias que contribuíram de forma significativa para o desenvolvimento, compartilhamento e controle das atividades desenvolvidas.

5.1.1 *Softwares* de Sistema de Informações Geográficas

Dois softwares SIG foram utilizados neste projeto, sendo eles o ArcGIS e o Quantum GIS (QGIS).

ArcGIS é um *software* SIG comercial desenvolvido pela empresa americana ERSI. O ArcGIS disponibiliza uma gama de ferramentas integradas e amigáveis (TANG & CLARK, 2003). Neste trabalho, foi utilizada a versão 10.8 do ArcMap para a criação dos mapas e seleção de dados.

O QGIS é uma solução *open-source* SIG, sendo um projeto oficial da *Open Source Geospatial Foundation* (OSGeo) e licenciado sob GPL (*General Public License*) (KURT MENKE *et al.*, 2016). O *software* teve sua primeira versão lançada em 2009 e, neste trabalho, adotou-se a versão 3.4 para a seleção, tratamento e unificação dos *rasters*.

5.1.2 Linguagens de Programação

Duas linguagens de programação foram utilizadas neste projeto, sendo elas Python e Julia.

O Python²⁴ é uma linguagem de programação interpretada, de alto nível e de propósito geral desenvolvida no início da década de 1990 (VAN ROSSUM, 2007). Nesse sentido, por ser uma das tecnologias mais difundidas nas áreas de Computação Científica e Aprendizado de Máquina (RASCHKA *et al.*, 2020), optou-se por utilizar o Python versão 3.6.10 para o desenvolvimento das etapas de análise exploratória²⁵, modelagem dos dados²⁶ e interpretação

²⁴ <https://www.python.org/>

²⁵ https://github.com/fnaghetini/Mapa-Preditivo/blob/main/1-exploratory_data_analysis.ipynb

²⁶ https://github.com/fnaghetini/Mapa-Preditivo/blob/main/2-predictive_litho_map.ipynb

do modelo²⁷. As informações acerca das bibliotecas do Python utilizadas neste trabalho são sintetizadas na Tabela 8.

Tabela 8 - Informações sobre as bibliotecas do Python utilizadas neste projeto.

Biblioteca	Versão	Descrição
Numpy	1.19.5	Manipulação de <i>arrays</i> n-dimensionais.
Pandas	1.0.5	Manipulação de <i>dataframes</i> ²⁸ .
Matplotlib	3.2.2	Visualização dos dados.
Seaborn	0.10.1	Visualização dos dados.
Scikit-Learn	0.24.1	Algoritmos de pré-processamento e Aprendizado de Máquina.
Imbalanced-Learn	0.8.0	Algoritmos e <i>pipelines</i> para lidar com classes desbalanceadas.
Xgboost	1.4.0	Algoritmo <i>XGBoost</i> .
Geopandas	0.9.0	Manipulação de <i>dataframes</i> geoespaciais.
Rasterio	1.1.7	Manipulação de <i>rasters</i> .
Folium	0.12.0	Visualização interativa de mapas.
Shap	0.39.0	<i>Framework</i> para explicação de modelos de Aprendizado de Máquina.

Julia²⁹ é uma linguagem de programação flexível e dinâmica com ênfase nas áreas de Computação Científica e Numérica e de performance comparável a linguagens estáticas tradicionais. Essa linguagem de programação, desenvolvida em 2012, é amigável como o Python e, ao mesmo tempo, de performance similar à linguagem C (BEZANSON *et al.*, 2012). A linguagem Julia versão 1.6.3 foi utilizada durante a análise de fenômenos característicos de dados geoespaciais, uma vez que conta com o GeoStats.jl³⁰, um *framework* geoestatístico de alta performance e *open-source*.

5.1.3 Ambientes de Desenvolvimento

Neste trabalho, foram utilizados três ambientes de desenvolvimento, sendo eles Jupyter Notebook, Pluto e Visual Studio Code.

O Jupyter Notebook é uma plataforma *open-source* interativa capaz de executar códigos via navegador (RANDLES *et al.*, 2017; KLUYVER *et al.*, 2016). Esse ambiente gera documentos denominados *notebooks* que, por sua vez, são constituídos por células de código e texto que podem ser modificadas e executadas individualmente (KLUYVER *et al.*, 2016).

²⁷ https://github.com/fnaghetini/Mapa-Preditivo/blob/main/3-model_explanation.ipynb

²⁸ Estruturas tabulares.

²⁹ <https://docs.julialang.org>

³⁰ <https://github.com/JuliaEarth/GeoStats.jl>

Nesse sentido, como os *notebooks* permitem um registro detalhado de todo o fluxo de trabalho desenvolvido em um projeto de pesquisa, optou-se pela utilização do Jupyter Notebook como principal ambiente de desenvolvimento do presente projeto.

O Pluto³¹ é, assim como o Jupyter Notebook, uma plataforma interativa capaz de executar códigos pelo navegador e gerar *notebooks*. Os *notebooks* Pluto são compatíveis com a linguagem Julia e, além de serem interativos, são responsivos, uma vez que, quando uma variável ou função é modificada, todas as células dependentes são automaticamente atualizadas. Apenas o último *notebook*³² foi desenvolvido em linguagem Julia no ambiente Pluto.

O Visual Studio Code³³ é uma plataforma de código aberto da Microsoft que fornece recursos para edição de códigos em múltiplas linguagens de programação, como Python (MICROSOFT INC, 2021). Essa tecnologia foi adotada no desenvolvimento das funções auxiliares³⁴ que compõem o projeto.

5.1.4 Sistema de Versionamento

O Git é um sistema de controle de versão *open-source* que permite o registro de todas as modificações realizadas em um ou mais arquivos ao longo do tempo (CHACON & STRAUB, 2014). Sendo assim, essa tecnologia foi adotada como uma forma de rastrear as alterações realizadas no projeto desde o seu início, de modo que todas as versões criadas podem ser facilmente acessadas a qualquer momento.

O GitHub consiste em uma plataforma social aberta de compartilhamento de código que utiliza o Git como sistema de controle de versão (THUNG *et al.*, 2013). A possibilidade de publicação de projetos e pesquisas em repositórios públicos estimula a interação entre membros da comunidade e favorece o desenvolvimento colaborativo do conhecimento. Visando reproduzibilidade, tanto o banco de dados quanto todo o fluxo de trabalho desenvolvido no presente projeto encontram-se hospedados publicamente nessa plataforma³⁵.

³¹ <https://github.com/fonsp/Pluto.jl>

³² https://github.com/fnaghetini/Mapa-Preditivo/blob/main/4-geospatial_issues.jl

³³ <https://code.visualstudio.com/>

³⁴ <https://github.com/fnaghetini/Mapa-Preditivo/tree/main/functions>

³⁵ <https://github.com/fnaghetini/Mapa-Preditivo>

5.2 FLUXO DE TRABALHO

Nesta seção, cada uma das etapas realizadas é descrita detalhadamente. A Figura 12 ilustra o fluxo de trabalho completo do projeto.

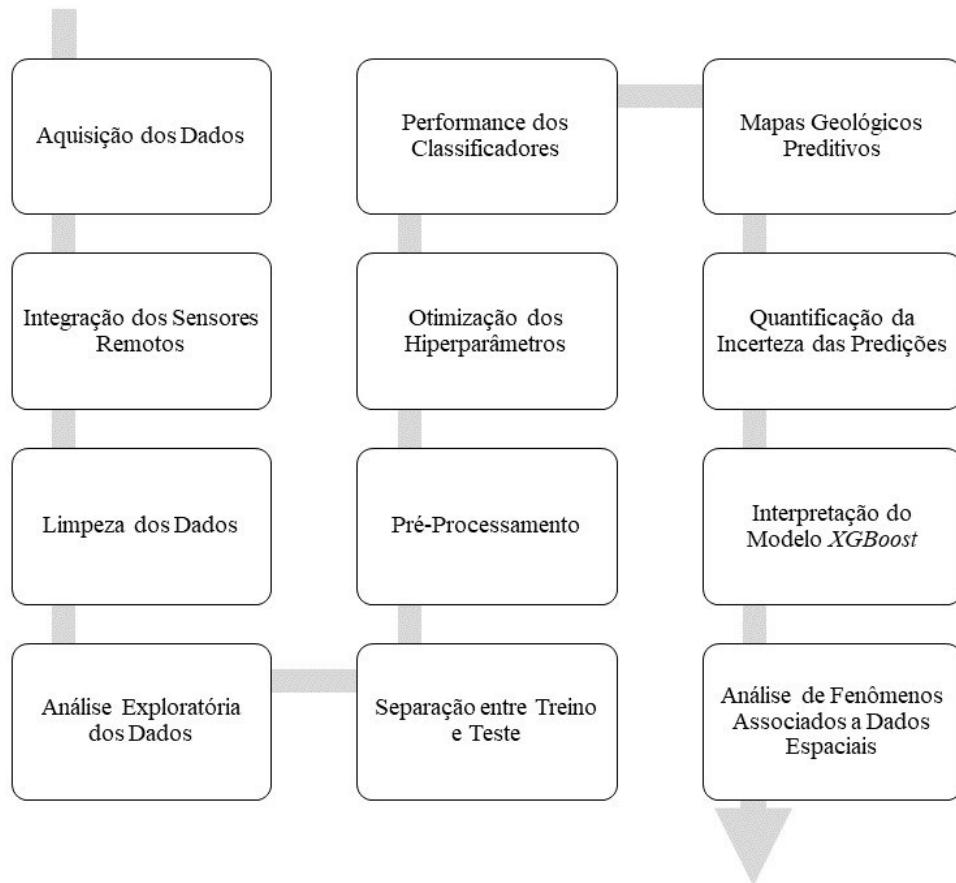


Figura 12 - Fluxo de trabalho do projeto.

5.2.1 Aquisição dos Dados

Os dados utilizados neste trabalho foram obtidos a partir de três fontes distintas. O mapa geológico integrado da região de Diamantina (Figura 11), confeccionado pelos discentes do curso de Geologia da UFMG, foi utilizado como fonte para a extração dos dados de unidades litoestratigráficas³⁶. Ressalta-se que algumas simplificações foram realizadas para atender os objetivos do trabalho, tais como a unificação dos níveis informais das formações São João da Chapada e Sopa-Brumadinho e a omissão das suítes metaígneas³⁷ Pedro Lessa e Conceição do Mato Dentro. Tanto essas modificações quanto a vetorização do mapa foram realizadas no software ArcGIS.

³⁶ A variável resposta deste trabalho.

³⁷ As suítes metaígneas apresentam ocorrência muito restrita na área (< 1%).

Apesar de se ter acesso aos pontos levantados em campo pelos alunos, optou-se por um processo de amostragem regular de pontos³⁸ diretamente do mapa integrado. Segundo KUHN *et al.* (2019), a amostragem de dados a partir de mapas geológicos maduros permite um posterior refinamento do mapeamento, além de possibilitar uma descrição quantitativa da incerteza associada à classificação das unidades litoestratigráficas. Por outro lado, a utilização de pontos de campo, normalmente agrupados preferencialmente em regiões de mais fácil acesso, resulta em modelos preditivos de baixa performance.

O segundo grupo de dados se refere às bandas Landsat 8 adquiridas no portal *Earth Explorer* da USGS³⁹. Os dados foram levantados pelo sensor no dia 16 de novembro de 2020⁴⁰ e adquiridos em formato *raster* (.tif).

Os dados geofísicos e o MDT, por sua vez, foram adquiridos via Biblioteca Vitória Pedersoli do Instituto de Geociências da UFMG em CD-ROM, no formato *grid* (.grd).

5.2.2 Integração dos Sensores Remotos

Os sensores remotos utilizados neste projeto apresentam sistemas de referência, resoluções e extensões distintos. Portanto, o objetivo desta etapa consiste na padronização e unificação desses sensores em uma estrutura tabular única $n \times p$, em que n representa as instâncias e p os atributos. A Figura 13 ilustra o fluxo de trabalho seguido para a preparação dos sensores remotos que, por sua vez, foi executado no *software* QGIS.

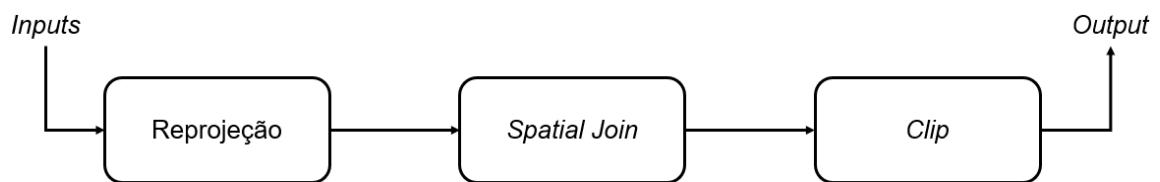


Figura 13 - Fluxo de trabalho executado para a integração dos sensores remotos.

Inicialmente, os sensores remotos e o mapa geológico foram reprojetados para um sistema de referência comum SIRGAS2000 UTM zona 23S via interpolação bilinear. Essa técnica, amplamente utilizada, consiste na interpolação linear ao longo dos eixos X e Y, respectivamente (KIM *et al.*, 2019).

³⁸ Com um espaçamento de 62,5 m entre os pontos.

³⁹ <https://earthexplorer.usgs.gov>

⁴⁰ Número de identificação (ID): LC08_L2SP_218072_20201116_20210315_02_T1.

Em seguida, todos os *inputs* foram agregados em uma única estrutura tabular a partir da tarefa *Spatial Join*. Esse procedimento objetiva combinar dois conjuntos de dados regionalizados a partir de relações espaciais entre ambos (ZHOU *et al.*, 1998). As dimensões de célula obtidas (62,5 m x 62,5 m) correspondem à resolução apresentada pelos *grids* geofísicos que, por sua vez, é a menor dentre os sensores remotos utilizados.

Por fim, os dados unificados foram balizados pelo polígono da área do projeto, a partir do processo *Clip*. O *output*⁴¹ obtido consiste em uma estrutura tabular com 11.418 instâncias e 17 atributos (Tabela 9). Esses dados foram utilizados como entrada nas etapas a seguir que, por sua vez, foram realizadas no ambiente Jupyter Notebook. Ressalta-se que as coordenadas espaciais (*i.e.* X e Y) não foram consideradas como variáveis independentes e, portanto, não participaram do treinamento dos algoritmos. Essas coordenadas têm, essencialmente, o papel de indexação espacial dos dados.

Tabela 9 - Dados resultantes do processo de integração dos sensores remotos.

X	Y	GT	K	Th	U	CT	U/K	Th/K	U/Th	MDT	B02	B03	B04	B06	B07	Target
$x_*^{(1)}$	$x_*^{(1)}$	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$	$x_4^{(1)}$	$x_5^{(1)}$	$x_6^{(1)}$	$x_7^{(1)}$	$x_8^{(1)}$	$x_9^{(1)}$	$x_{10}^{(1)}$	$x_{11}^{(1)}$	$x_{12}^{(1)}$	$x_{13}^{(1)}$	$x_{14}^{(1)}$	$y^{(1)}$
$x_*^{(2)}$	$x_*^{(2)}$	$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$	$x_4^{(2)}$	$x_5^{(2)}$	$x_6^{(2)}$	$x_7^{(2)}$	$x_8^{(2)}$	$x_9^{(2)}$	$x_{10}^{(2)}$	$x_{11}^{(2)}$	$x_{12}^{(2)}$	$x_{13}^{(2)}$	$x_{14}^{(2)}$	$y^{(2)}$
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	
...	...	$x_1^{(11418)}$	$y^{(11418)}$

5.2.3 Limpeza dos Dados

Um procedimento de truncamento dos canais radiométricos (*i.e.* U, Th e K) foi realizado com o intuito de tratar amostras com valores extremos, sejam eles muito baixos (por vezes negativos) ou muito altos.

Nesse sentido, para a identificação desses valores anômalos, foram calculados um limiar inferior (LI) e um limiar superior (LS). O LI foi estabelecido como um décimo da média da variável radiométrica \bar{X} :

$$LI = \bar{X}/10 \quad (5.1)$$

Por outro lado, o LS foi calculado a partir do percentil 99,5 do canal radiométrico $P99,5(X)$:

⁴¹ <https://github.com/fnaghetini/Mapa-Preditivo/tree/main/data>

$$LS = P99,5(X) \quad (5.2)$$

Em seguida, todos os valores iguais ou inferiores ao LI foram igualados ao valor de LI. De forma análoga, todos os valores iguais ou superiores ao LS foram substituídos pelo próprio valor de LS.

5.2.4 Análise Exploratória dos Dados

A Análise Exploratória dos Dados (AED) visa aplicar técnicas estatísticas para sumarizar e obter *insights* a partir da descrição de um conjunto de dados. Segundo TUKEY (1977), a AED é análoga a um trabalho de investigação, já que, para realizá-la, são necessários ferramentas e conhecimento por parte do analista.

Neste trabalho, a AED foi executada essencialmente para descrever e sumarizar o comportamento das variáveis independentes e seu relacionamento com a variável dependente. Na descrição univariada, gráficos (*e.g.* histogramas e *boxplots*) e estatísticas (*e.g.* média, coeficiente de variação e coeficiente de assimetria) foram utilizados com o intuito de compreender a distribuição das variáveis independentes quando agrupadas pelas unidades litoestratigráficas. Já na descrição bivariada, o objetivo foi investigar a correlação (*no feature space*) entre as variáveis independentes a partir de gráficos (*e.g.* matriz de correlação linear) e estatísticas (*e.g.* coeficiente de correlação linear).

5.2.5 Separação entre Treino e Teste

Segundo CRACKNELL & READING (2014), os dados de treino (T_a) são utilizados para treinar algoritmos e estimar a performance dos modelos gerados. Por outro lado, os dados de teste (T_b) são isolados e, portanto, não participam da etapa de aprendizado dos modelos (GUYON, 2009). Nesse sentido, o conjunto T_b , por não ser visto pelo modelo, é utilizado para avaliar sua performance sem que haja viés (CRACKNELL & READING, 2014).

Diversas estratégias podem ser adotadas para a construção do conjunto de treino. CRACKNELL *et al.* (2014), KUHN *et al.* (2019) e COSTA *et al.* (2019), por exemplo, definiram o conjunto de treino a partir da amostragem aleatória de 100 exemplos por unidade geológica. LEVERINGTON (2010), por outro lado, amostrou aleatoriamente 300 instâncias por classe, exceto no caso de uma unidade gabroica de ocorrência restrita, em que apenas 60 observações foram reamostradas.

Neste trabalho, a heurística adotada foi a definição do conjunto T_a a partir da amostragem aleatória de 150 exemplos por unidade litoestratigráfica. No caso do Complexo Granito-Gnáissico que, por sua vez, apresenta apenas 67 instâncias, 70% delas foram aleatoriamente amostradas para também compor T_a . As demais amostras, por outro lado, foram selecionadas para compor o conjunto T_b .

5.2.6 Pré-Processamento

O pré-processamento dos dados é uma etapa crucial na construção de modelos de Aprendizado de Máquina. Ele consiste na adequação, substituição e criação de novas *features* a partir dos dados iniciais (HEATON, 2016; THAKUR, 2020). Esta etapa exige criatividade e experiência do analista com os dados para aplicação de certas técnicas (*e.g.* codificação de variáveis, estandardização, normalização, análise de componentes principais), uma vez que cada modelo responde de forma diferente aos passos de pré-processamento adotados (HEATON, 2016).

Neste trabalho, as seguintes técnicas de pré-processamento foram aplicadas em sequência:

- i. Escalonamento das *features*;
- ii. Redução da dimensionalidade;
- iii. Superamostragem.

Após definidas, essas etapas foram encadeadas para a construção de um *pipeline*⁴², visando automatizar toda a rotina (Figura 14).

⁴² Objeto que armazena uma sequência linear de etapas.

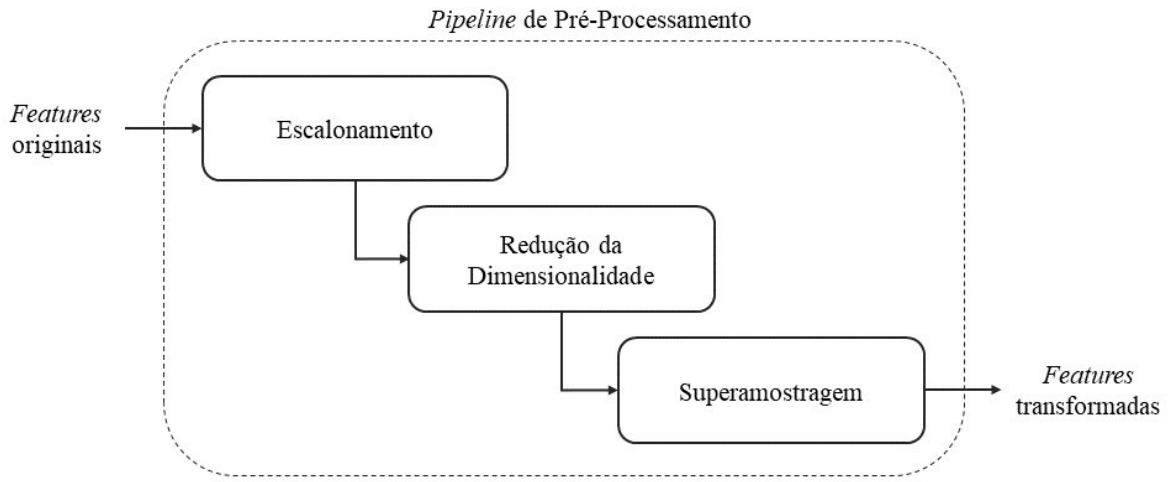


Figura 14 - Pipeline de pré-processamento dos dados adotado no trabalho.

5.2.6.1 Escalonamento das *Features*

O escalonamento das *features* tem como objetivo principal a padronização da escala dos dados (KUHN & JOHNSON, 2019). Esse método, por vezes, é essencial quando se utiliza algoritmos lineares, como SVC e RL (THAKUR, 2020).

A escolha do método de escalonamento apropriado depende da necessidade do analista, uma vez que cada um deles possui um objetivo distinto. A estandardização (*standardization*), técnica adotada neste trabalho, consiste na transformação das variáveis preditoras originais X em variáveis Z que seguem uma distribuição padrão, com média igual a 0 e desvio padrão igual a 1. Essa transformação, para uma determinada *feature* com valores x , média μ e desvio padrão σ , é calculada como:

$$z = \frac{x - \mu}{\sigma} \quad (5.3)$$

tal que que $z \in Z$ e $Z \sim N(0,1)$.

5.2.6.2 Redução da Dimensionalidade

O termo “Maldição da Dimensionalidade” (*Curse of Dimensionality*), introduzido por Bellman na década de 1950, descreve o problema causado pelo aumento exponencial do volume associado à adição de novas dimensões no espaço euclidiano (KEOGH & MUEEN, 2017). Uma consequência desse fenômeno é que, para se obter uma estimativa razoável de uma função f , o número de instâncias necessárias cresce exponencialmente com o número de *features* (*i.e.* dimensões) adicionadas.

A redução da dimensionalidade apresenta-se como uma alternativa viável para mitigar o problema descrito acima. Essa estratégia visa reduzir o número de dimensões de um conjunto

de dados, sem que haja uma perda expressiva de informações (VAN DER MAATEN *et al.*, 2009). A aplicação de técnicas de redução da dimensionalidade viabiliza não somente a visualização de dados, mas principalmente a eliminação de informações redundantes (MEDEIROS & COSTA, 2008).

A Análise de Componentes Principais (ACP), técnica utilizada neste trabalho, consiste em um algoritmo que objetiva encontrar projeções \tilde{x}_n similares às instâncias originais x_n , mas que apresentam uma dimensionalidade significativamente menor (DEISENROTH *et al.*, 2020). A partir da ACP, são calculadas as componentes principais (Figura 15), ou seja, vetores ortogonais entre si no *feature space* que definem um novo sistema de coordenadas em que os dados originais serão projetados (BURKOV, 2019). Nesse sentido, é comum que apenas as primeiras componentes principais sejam utilizadas, uma vez que elas tendem a preservar grande parte da variação original dos dados. Ressalta-se que apenas as cinco bandas Landsat 8 foram submetidas à ACP.

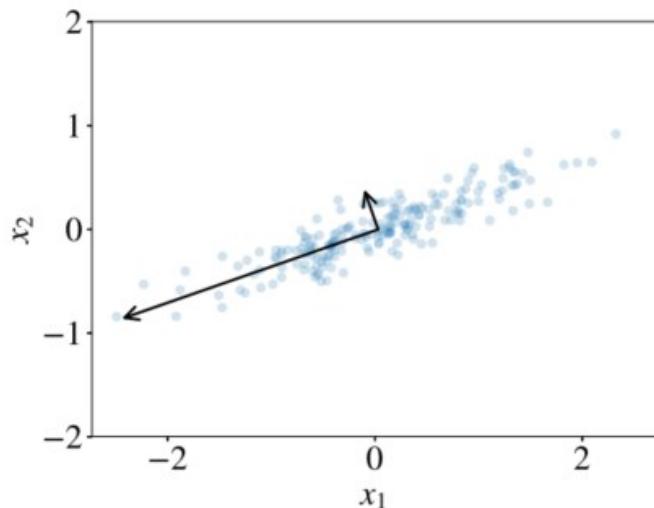


Figura 15 - Exemplo de ACP, em que os vetores representam as componentes principais (BURKOV, 2019).

5.2.6.3 Superamostragem

Dados desbalanceados ocorrem quando há uma diferença significativa entre as frequências (número de amostras) das classes de um determinado problema (SANTOS *et al.*, 2018). Esse contexto afeta diretamente a qualidade do modelo construído, uma vez que as unidades minoritárias, ou seja, aquelas que possuem menos instâncias, serão classificadas de forma equivocada pelo modelo (LAURIKKALA, 2001).

A superamostragem (*oversampling*) é uma, dentre diversas estratégias, que pode ser adotada para contornar o problema do desbalanceamento de classes a partir da geração de

novos exemplos das classes minoritárias (BURKOV, 2019). De forma geral, modelos construídos a partir de dados balanceados tendem a apresentar melhores performances (WEISS & PROVOST, 2001; LAURIKKALA, 2001). Diversas técnicas de superamostragem foram desenvolvidas, tais como *Random Oversampling*, SMOTE (*Synthetic Minority Oversampling Technique*) (CHAWLA *et al.*, 2002), ADASYN (*Adaptive Synthetic Sampling Technique*) (HE *et al.*, 2008), CBSO (*Cluster-Based Synthetic Oversampling*) (BARUA *et al.*, 2011).

Neste trabalho, foi adotada a técnica SMOTE para lidar com a situação de desbalanceamento das classes. Essa técnica consiste na criação de amostras sintéticas da classe minoritária a partir de operações realizadas no *feature space*, ao invés da simples superamostragem com reposição da classe com menor número de observações (CHAWLA *et al.*, 2002). Diferentemente do *Random Oversampling* que replica instâncias até que o balanceamento de classes seja atingido, o SMOTE, por gerar novas amostras, tende a evitar o sobreajuste do modelo (SANTOS *et al.*, 2018). Dada uma instância da classe minoritária x , seleciona-se os k exemplos mais próximos a x no *feature space* e os armazena no subconjunto S_k . Um novo exemplo sintético x' é então gerado a partir da seguinte operação:

$$x' = x + \lambda(x_s - x) \quad (5.4)$$

em que $\lambda \in [0,1]$ é o hiperparâmetro de interpolação aleatoriamente escolhido e x_s é uma instância aleatoriamente amostrada do subconjunto S_k (BURKOV, 2019).

5.2.7 Otimização dos Hiperparâmetros

Em geral, a construção de modelos de Aprendizado de Máquina é complexa e demanda muito tempo, visto que, por serem genéricos, diferentes algoritmos tendem a ser mais adequados a diferentes problemas e bases de dados (ZÖLLER & HUBER, 2021). Além dos parâmetros do modelo, que são atualizados durante a etapa de aprendizado, existem ainda os hiperparâmetros que, por não serem diretamente estimados a partir dos dados, devem ser definidos antes da etapa de treinamento e influenciam diretamente a performance dos modelos (KUHN & JOHNSON, 2013). O processo em que a arquitetura ideal do modelo é definida a partir da escolha dos hiperparâmetros ótimos é denominado otimização dos hiperparâmetros⁴³ (YANG & SHAMI, 2020).

⁴³ Também chamado de *tuning* dos hiperparâmetros.

Neste trabalho, duas estratégias de otimização dos hiperparâmetros foram adotadas *Grid Search* e *Random Search*. Em ambos os casos, um *grid* de valores é definido como o espaço de busca inicial e múltiplas combinações desses valores são avaliadas de acordo com alguma métrica de referência que, neste caso, é o *F1-score*. Por fim, a combinação de hiperparâmetros que fornece a melhor performance é selecionada. Os valores selecionados são denominados hiperparâmetros ótimos.

No *Grid Search*, todas as possíveis combinações entre os valores do *grid* inicial são avaliadas, o que faz com que esse método tenha um elevado custo computacional. Nesse sentido, essa técnica foi adotada durante a otimização dos hiperparâmetros dos modelos mais simples⁴⁴ (*i.e.* Regressão Logística, *Decision Tree*, *Naive Bayes*, *K-Nearest Neighbors* e *Support Vector Machine*), considerando uma validação cruzada com cinco *folds*.

Já no caso do *Random Search*, são selecionadas combinações aleatórias de valores a partir do espaço de busca inicial. Entretanto, como se define um número fixo de combinações, esse método pode apresentar um custo computacional inferior ao do *Grid Search*. Dessa forma, o *Random Search* foi utilizado apenas para o *tuning* dos hiperparâmetros dos classificadores mais complexos⁴⁵ (*i.e.* *Random Forest*, *XGBoost* e *Multilayer Perceptron*), também utilizando uma validação cruzada com cinco *folds*.

A Tabela 10 apresenta o *grid* de valores avaliados durante a otimização dos hiperparâmetros para cada um dos oito modelos pré-selecionados.

⁴⁴ Menor número de hiperparâmetros.

⁴⁵ Maior número de hiperparâmetros.

Tabela 10 – Grid de valores avaliados durante a otimização dos hiperparâmetros. Os nomes dos hiperparâmetros seguem a documentação do framework Scikit-Learn⁴⁶.

Algoritmos	Hiperparâmetros	Valores	Algoritmos	Hiperparâmetros	Valores
RL	<i>solver</i>	newton-cg; saga; sag; lbfgs	KNN	<i>n_neighbors</i>	3; 5; 7; 9; 11; 13; 15; 17; 19; 21; 23; 25; 27; 29
	<i>C</i>	1E-03; 4,64E-03; 2,15E-02; 1E-01; 4,64E-01; 2,15; 10,00; 4,64E+01; 2,15E+02; 1.00E+03		<i>weights</i>	uniform; distance
SVM	<i>C</i>	1E-03; 4,64E-03; 2,15E-02; 1E-01; 4,64E-01; 2,15; 10,00; 4,64E+01; 2,15E+02; 1.00E+03		<i>p</i>	1; 2
	<i>gamma</i>	0,001; 0,01; 0,1; 1,0; 10,0; 100,0; auto		<i>n_estimators</i>	25; 50; 100; 500
NB	<i>kernel</i>	poly; rbf	RF	<i>max_depth</i>	15; 25; 30; none
	<i>var_smoothing</i>	1E-10; 1E-09; 1E-08; 1E-07; 1E-06; 1E-05; 1E-04; 1E-03; 1E-02; 1E-01; 1,0		<i>criterion</i>	gini; entropy
DT	<i>criterion</i>	gini; entropy		<i>min_samples_split</i>	1; 2; 5; 10
	<i>max_depth</i>	15; 25; 30; none		<i>min_samples_leaf</i>	1; 2; 5; 10
MLP	<i>hidden_layer_sizes</i>	(10); (20); (30); (50); (10, 10); (20, 20); (10, 10, 10); (10, 30, 10); (20, 20, 20)	XGB	<i>eta</i>	0,01; 0,015; 0,025; 0,05; 0,1
	<i>activation</i>	logistic; tahn; relu		<i>gamma</i>	0,05; 0,10; 0,30; 0,40; 0,50; 0,70; 0,90; 1
	<i>solver</i>	lbfgs; sgd; adam		<i>max_depth</i>	3; 5; 7; 9; 12; 15; 17; 25
	<i>alpha</i>	0,001; 0,01; 0,1; 1; 10; 100; 1000		<i>subsample</i>	0,6; 0,7; 0,8; 0,9; 1,0
	<i>learning_rate</i>	constant; adaptive		<i>colsample_bytree</i>	0,6; 0,7; 0,8; 0,9; 1,0
	<i>learning_rate_init</i>	0,0001; 0,001; 0,01; 0,1; 0,15; 0,20; 0,25; 0,30		<i>min_child_weight</i>	1; 3; 5; 7
	<i>max_iter</i>	1; 5; 10; 20; 50; 80; 100; 150; 200; 250		<i>reg_lambda</i>	0,001; 0,01; 0,1; 1; 10; 100; 1000
				<i>alpha</i>	0,001; 0,01; 0,1; 1,0; 10; 100; 1000

5.2.8 Performance dos Classificadores

A matriz de confusão é uma estrutura tabular $C \times C$ comumente utilizada em problemas de classificação como uma forma de visualizar a performance dos modelos, sendo C o número de classes (VISA *et al.*, 2011). Cada linha da matriz representa as instâncias pertencentes a uma classe real⁴⁷, enquanto cada coluna representa as previsões de uma determinada classe.

Cada célula da matriz de confusão pode ser caracterizada como verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) ou falso negativo (FN). Considerando uma classe c de referência, VP se refere ao número de instâncias corretamente classificadas como classe c . VN diz respeito ao número de exemplos que não foram preditos como classe c e de fato não pertencem a essa classe. FP indica o número de observações que foram classificadas como classe c quando na realidade não pertencem a essa classe. Por fim, FN se

⁴⁶ <https://scikit-learn.org/0.24/index.html>

⁴⁷ Neste trabalho, a classe real é aquela associada ao mapa geológico integrado de campo.

refere ao número de instâncias que não foram preditas como classe c , quando na verdade pertencem a essa classe. A Figura 16 mostra uma matriz de confusão 6×6 , em que c_2 é a classe avaliada.

		Predito					
		c_1	$* c_2$	c_3	c_4	c_5	c_6
Real	c_1	VN	FP	VN	VN	VN	VN
	$* c_2$	FN	VP	FN	FN	FN	FN
	c_3	VN	FP	VN	VN	VN	VN
	c_4	VN	FP	VN	VN	VN	VN
	c_5	VN	FP	VN	VN	VN	VN
	c_6	VN	FP	VN	VN	VN	VN

- Predições corretas
- Predições incorretas
- * Classe avaliada

Figura 16 – Exemplo de matriz de confusão 6×6 da classe c_2 . VP = verdadeiro positivo, VN = verdadeiro negativo, FN = falso negativo, FP = falso positivo. Figura elaborada pelos autores.

Quatro métricas de performance comumente utilizadas em problemas de classificação foram escolhidas, sendo elas acurácia, precisão, *recall* e *F1-score*. Ressalta-se que *F1-score* foi selecionada como métrica principal, por sumarizar os *scores* de precisão e *recall* e por ser mais adequada em problemas de classes desbalanceadas do que a acurácia. Essas medidas são calculadas individualmente para cada classe a partir da matriz de confusão e suas equações são apresentadas abaixo:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (5.5)$$

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (5.6)$$

$$\text{Recall} = \frac{VP}{VP + FN} \quad (5.7)$$

$$F1 - score = \frac{2(precisão)(recall)}{precisão + recall} \quad (5.8)$$

Ressalta-se que os valores de precisão, *recall* e *F1-score* apresentados são resultantes da média dessas métricas entre todas as unidades ponderada pelo número de exemplos pertencentes a cada classe. Portanto, *scores* associados a classes mais frequentes recebem pesos maiores do que *scores* de unidades mais restritas.

As estimativas das métricas de performance acima foram obtidas a partir da técnica VCKF com cinco *folds*. É importante ressaltar que, exclusivamente durante essa estimação, as etapas de pré-processamento foram realizadas durante a VCKF⁴⁸. Em outras palavras, essas etapas foram aplicadas após a separação dos cinco *folds*, de modo que, a cada iteração, apenas os dados pertencentes aos *folds* de treino foram pré-processados (Figura 17). Essa abordagem foi seguida, já que, segundo SANTOS *et al.* (2018), caso a VCKF fosse implementada antes da superamostragem, padrões similares poderiam ser encontrados nos *folds* de treino e validação e, consequentemente, estimativas de performance superotimistas seriam geradas.

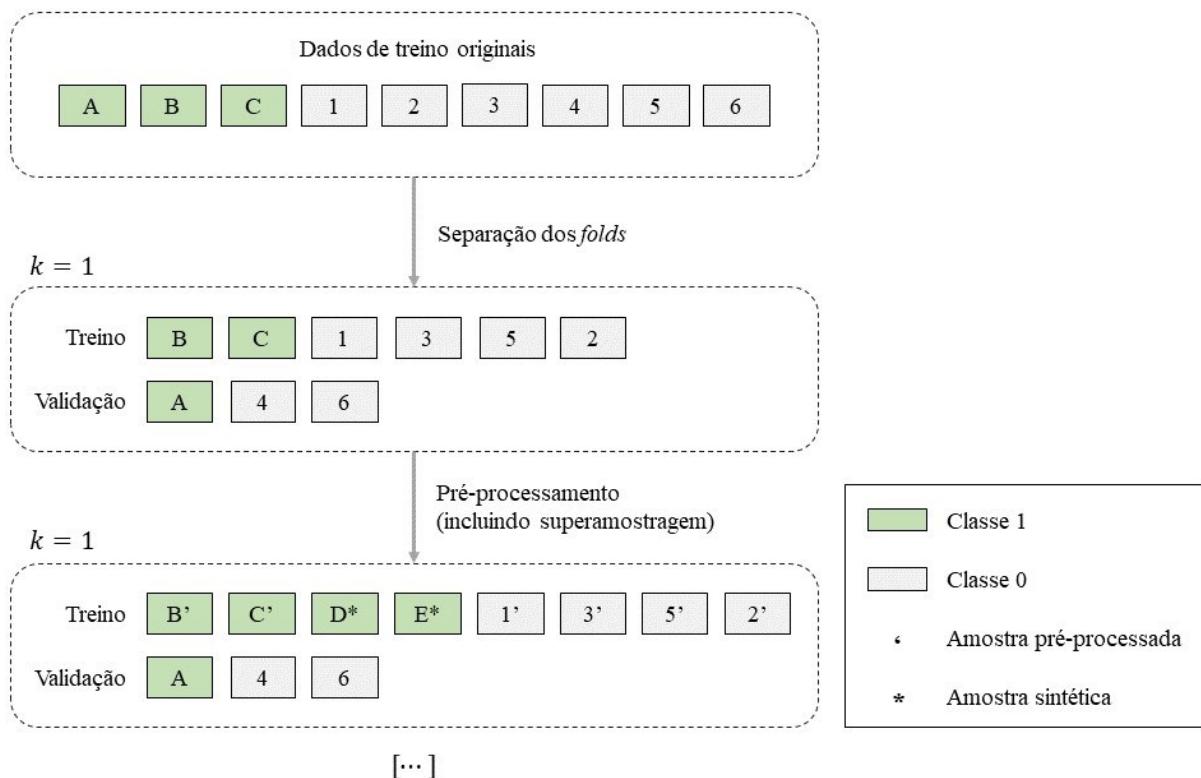


Figura 17 - Superamostragem durante a VCKF em um problema de classificação binária. Apenas o processo na primeira iteração ($k=1$) é apresentado. Figura elaborada pelos autores e baseada em SANTOS *et al.* (2018).

⁴⁸ A biblioteca *Imbalanced-Learn* do Python fornece essa possibilidade.

Além disso, a performance dos classificadores também foi avaliada no conjunto de teste. Nesse caso, as mesmas quatro métricas foram consideradas, sendo elas também ponderadas pelo número de instâncias de cada classe.

5.2.9 Mapas Geológicos Preditivos

A elaboração e tratamento dos mapas foram realizadas no ambiente Jupyter Notebook⁴⁹ e no software ArcMap. Ao todo, foram confeccionados nove mapas preditivos, de modo que oito deles representam as previsões geradas por cada modelo. Essas previsões foram então exportadas como *rasters*⁵⁰ e tratadas no ArcMap, onde os mapas finais foram elaborados.

O mapa restante foi confeccionado no ArcMap, a partir da interpolação das previsões geradas pelo modelo *XGBoost*. Essas previsões também foram obtidas no ambiente Jupyter Notebook, mas nesse caso foram exportadas como pontos⁵¹. A interpolação foi realizada com a ferramenta *Point to Raster* que, por sua vez, converte pontos de entrada em um objeto *raster*. A variável interpolada contém valores discretos de 1 a 6 que representam cada uma das seis unidades litoestratigráficas. A resolução definida para os *pixels* do *raster* de saída foi de 62,5 metros, ou seja, o mesmo tamanho de célula dos sensores remotos.

No *raster* de saída do modelo *XGBoost*, foram ainda aplicadas técnicas de redução de ruídos. Considerando que um valor de classe é atribuído a cada célula do *raster*, a ferramenta *Majority Filter* foi utilizada de forma a substituir células do *raster* pelo valor da classe majoritária entre as k células vizinhas contíguas. Optou-se então por utilizar um valor de $k = 8$, além de uma condição de restrição em que ao menos 5 das 8 células contíguas devem pertencer à mesma classe para que a substituição ocorra. Por fim, uma segunda técnica, que consiste na reamostragem das classes durante a visualização do mapa, foi aplicada. Nesse caso, como os dados são discretos, a estratégia de reamostragem considerou o valor da classe majoritária entre os 6 *pixels* vizinhos.

Além dos mapas preditivos, foram também confeccionados oito mapas de inconsistências associados às previsões fornecidas por cada classificador. Esses mapas são constituídos por classificações corretas e incorretas. Uma previsão é considerada correta, caso o valor da classe predita por um modelo seja igual ao valor da unidade do mapa geológico

⁴⁹ https://github.com/fnaghetini/Mapa-Preditivo/blob/main/2-predictive_litho_map.ipynb

⁵⁰ <https://github.com/fnaghetini/Mapa-Preditivo/tree/main/output/rasters>

⁵¹ <https://github.com/fnaghetini/Mapa-Preditivo/blob/main/output/points/XGB.csv>

integrado (Figura 11). Por outro lado, uma classificação é incorreta, se a classe predita é diferente daquela observada no mapa integrado.

Foram ainda elaborados mapas de probabilidade por classe dos modelos *Random Forest* e *XGBoost*. Segundo KUHN *et al.*, (2019), esses mapas podem ser utilizados para avaliar a confiança das previsões realizadas pelo modelo para cada classe, individualmente. Nesse sentido, quanto maior é o valor de probabilidade predito para uma unidade, maior é a confiança dessa previsão. Como existem seis unidades litoestratigráficas na área de estudo, a previsão de um modelo é um vetor de probabilidades com seis elementos que, se somados, totalizam em 1. A Figura 18 ilustra um exemplo de vetor de probabilidade predito \hat{p} pelo modelo *Random Forest*.

$$\hat{p} = \begin{bmatrix} 0,65 \\ 0,18 \\ 0,10 \\ 0,07 \\ 0,00 \\ 0,00 \end{bmatrix} \longrightarrow \begin{array}{l} \text{Cx. Granito-Gnáissico} \\ \text{Fm. Barão de Guaicuí} \\ \text{Fm. Bandeirinha} \\ \text{Fm. São João da Chapada} \\ \text{Fm. Sopa-Brumadinho} \\ \text{Fm. Galho do Miguel} \end{array}$$

Figura 18 - Exemplo de um vetor de probabilidades predito para uma determinada instância. Figura elaborada pelos autores.

Ressalta-se que todos os mapas mencionados possuem uma resolução de 62,5 metros e encontram-se em SIRGAS 2000 UTM Zona 23S.

5.2.10 Quantificação da Incerteza das Previsões

Os valores de probabilidade podem ser visualizados espacialmente, para cada classe, a partir dos já mencionados mapas de probabilidade. Entretanto, a distribuição dessas probabilidades pode ser quantificada por um único número (KUHN *et al.*, 2019).

Neste estudo, optou-se pela utilização da entropia da informação como medida de quantificação de incerteza. Proposta por Shannon na década de 1940, essa medida já se provou ser eficiente na definição da distribuição espacial da incerteza (WELLMANN & REGENAUER-LIEB, 2012; KUHN *et al.*, 2016; KUHN *et al.*, 2019). A entropia da informação H , para uma instância, pode ser calculada como:

$$H = - \sum_{c=1}^C p_c \log p_c \quad (5.9)$$

em que p_c é a probabilidade predita para a c -ésima classe e C é o número total de classes. O mínimo valor que a entropia pode assumir é zero, quando $\log 1 = 0$. O logaritmo da equação (5.9) pode assumir diferentes bases, mas, neste trabalho, adotou-se a base 2, também utilizada por WELLMANN & REGENAUER-LIEB (2012). Ressalta-se que, para o cálculo da entropia, probabilidades nulas foram ignoradas.

De forma geral, quanto maior é o valor de entropia, maior é a incerteza associada às previsões. Por outro lado, quanto menor é a entropia calculada para uma instância, menor é a incerteza relacionada à sua previsão. Assim como os mapas de probabilidade por classe, os mapas de entropia da informação foram confeccionados apenas para os modelos *Random Forest* e *XGBoost*.

5.2.11 Interpretação do Modelo

A capacidade de interpretar as previsões geradas por um classificador é de suma importância. A interpretação de um modelo permite uma maior compreensão do processo que está sendo modelado, além de fornecer pistas sobre aspectos de melhoria do modelo (LUNDBERG & LEE, 2017).

Neste trabalho, a abordagem utilizada para interpretar as previsões geradas pelo modelo selecionado foi o *framework* SHAP. Para cada previsão, esse método baseado na teoria dos jogos atribui valores de importância às variáveis independentes. Segundo CHEN (2021), esses valores, denominados valores de Shapley, indicam o impacto (positivo ou negativo) das variáveis explicativas sobre uma determinada previsão e podem ser calculados a partir da equação (2.28).

As interpretações das previsões do modelo foram reportadas por meio dos chamados *summary plots*. Quando representado por um gráfico de barras empilhadas, o *summary plot* informa o impacto geral que cada variável preditora teve nas previsões das unidades litoestratigráficas. As variáveis são elencadas em ordem crescente de importância, de modo que o eixo horizontal representa a média dos módulos dos valores de Shapley, enquanto o eixo vertical contém as variáveis preditoras. A Figura 19 ilustra um exemplo de *summary plot*, em que a Variável 1 mostra-se a *feature* de maior importância, principalmente na previsão das classes 1, 3 e 4.

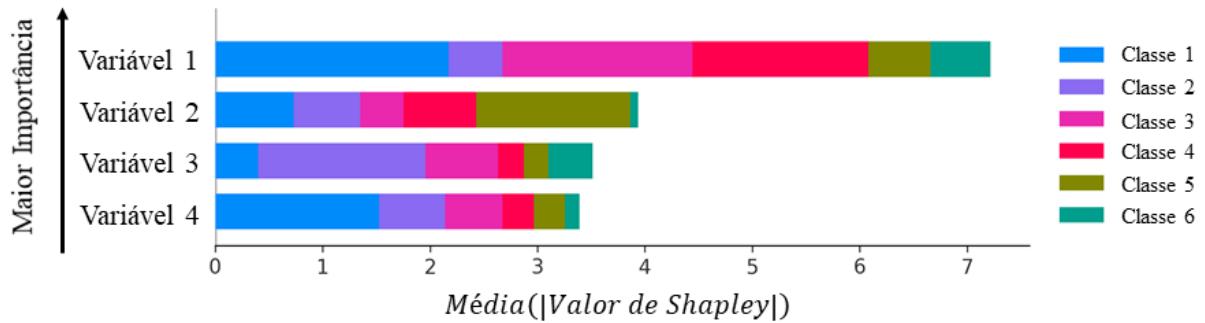


Figura 19 - Exemplo de summary plot que apresenta o impacto geral das features na predição das classes.
Figura elaborada pelos autores.

O *summary plot* pode ainda ser criado para cada classe individualmente e, nesse sentido, é representado por um *strip plot*⁵². Nesse caso, o *summary plot* informa o impacto (positivo ou negativo) que cada uma das variáveis independentes teve na predição de uma unidade, quando essas variáveis assumem valores altos e baixos. Nesse gráfico, as variáveis preditoras também são elencadas em ordem crescente de importância. A Figura 20 exibe um exemplo de *summary plot* para uma classe, em que valores elevados da Variável 1 (variável de maior importância) favorecem a predição dessa classe, ao passo que valores baixos levam o modelo a não predizê-la.

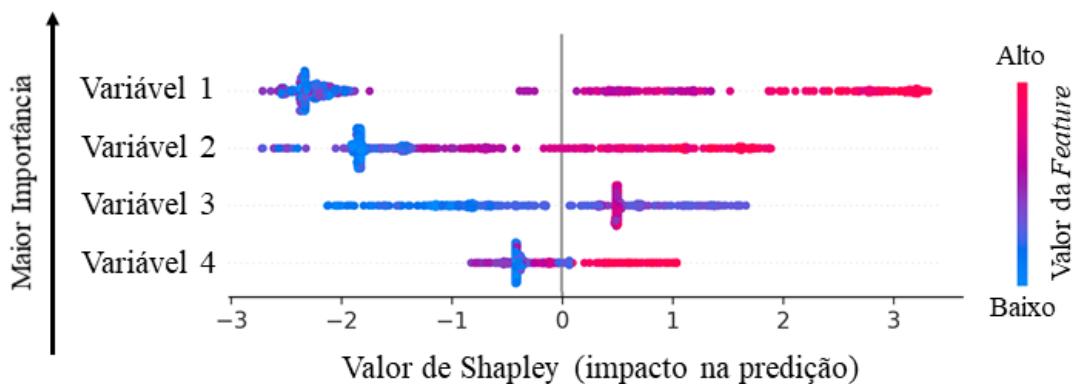


Figura 20 - Exemplo de summary plot para uma determinada classe. Figura elaborada pelos autores.

5.2.12 Análise de Fenômenos Associados a Dados Geoespaciais

Os algoritmos clássicos de Aprendizado de Máquina, em geral, modelam as observações com base em relações no *feature space* e, portanto, não consideram a presença de dependência espacial entre as VR (TALEBI *et al.*, 2021). Além disso, a utilização de técnicas clássicas de reamostragem (*e.g.* VCKF) em problemas geoespaciais pode levar a estimativas

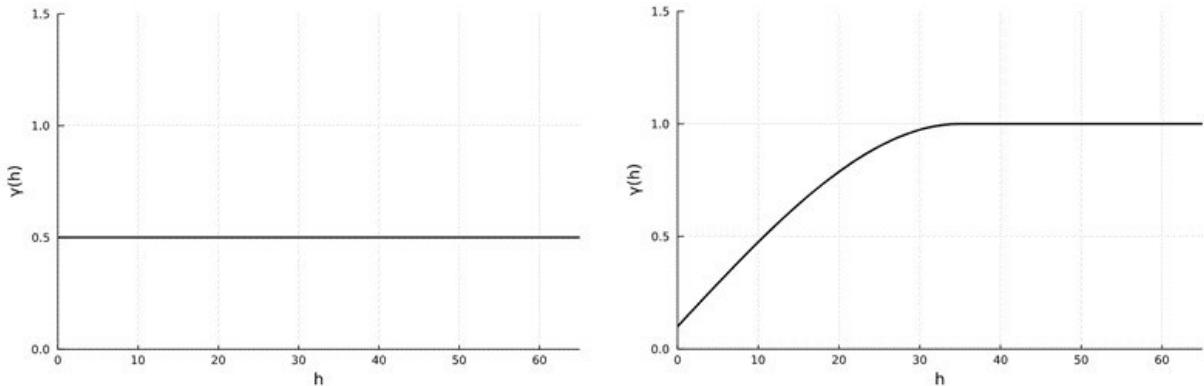
⁵² Diagrama de dispersão em que uma das variáveis é contínua e outra é categórica.

do erro de generalização inapropriadas em função da violação de algumas premissas (HOFFIMANN *et al.*, 2021).

Nesse sentido, algumas análises simples foram conduzidas com o intuito de se identificar a presença de fenômenos e propriedades tipicamente associados a dados geoespaciais. Uma dessas propriedades, a correlação espacial, pode ser definida como a similaridade entre valores de uma mesma variável, em função da proximidade entre amostras no espaço geográfico (GRIFFTH, 2003). Uma das estatísticas mais comuns para quantificar a dependência espacial entre pares de amostras, o variograma $\gamma(h)$, pode ser calculado como:

$$\gamma(h) = \frac{1}{2n} \sum_{i=1}^n [Z(u_i) - Z(u_i + h)]^2 \quad (5.10)$$

em que n é o número de pares de amostras e $Z(u_i) - Z(u_i + h)$ representa a diferença entre os valores assumidos por duas amostras separadas por um vetor h no espaço geográfico. Nesse sentido, uma variável apresenta correlação espacial quando o seu variograma possui um alcance positivo não negligenciável (HOFFIMANN *et al.*, 2021). A Figura 21 apresenta uma comparação entre o variograma de uma variável com ausência de correlação espacial (*i.e.* uma VA) e um variograma de uma variável com estrutura espacial (*i.e.* uma VR).



*Figura 21 - Exemplos de variogramas na ausência (esquerda) e presença (direita) de correlação espacial.
Figura elaborada pelos autores.*

Um outro fenômeno típico de dados geoespaciais é a distorção (*shift*) da distribuição bivariada entre os conjuntos de treino e teste. Segundo HOFFIMANN *et al.* (2021), esse fenômeno ocorre quando a função estimada no domínio de treino é distinta da função associada ao conjunto de teste. Uma forma simples de se avaliar a presença de *shift* na distribuição bivariada é a comparação entre diagramas de dispersão dos conjuntos de treino e teste. A Figura 22 ilustra uma situação em que há uma visível distorção na distribuição

bivariada das variáveis *gamma-ray* e potencial espontâneo entre os dados de treino e teste. Ressalta-se que, nesse caso, os conjuntos de treino e teste situam-se em áreas geograficamente distintas.

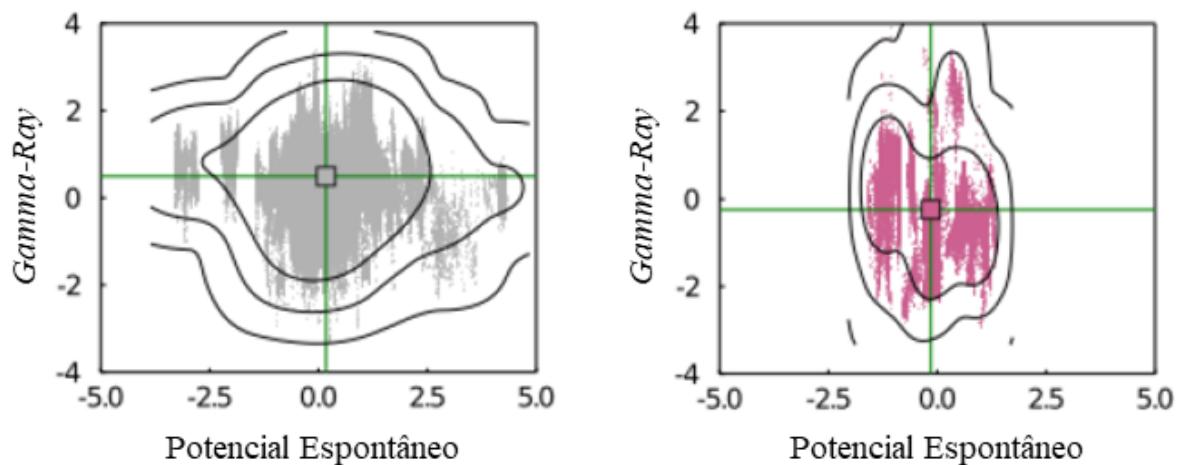


Figura 22 - Distorção na distribuição bivariada entre os dados utilizados para o treinamento do modelo (cinza) e os dados de teste (rosa). As variáveis geofísicas se encontram estandardizadas. Modificado de HOFFIMANN et al. (2021).

6 RESULTADOS

Nesta seção são apresentados os resultados obtidos durante o trabalho. Primeiramente, uma análise exploratória dos dados foi conduzida para descrever e sumarizar informações preliminares do conjunto de dados. Uma heurística de separação dos dados em conjuntos de treino e teste foi adotada, de modo que as amostras do primeiro conjunto foram, em seguida, submetidas às etapas de pré-processamento. Os hiperparâmetros dos oito algoritmos avaliados foram então otimizados e os *scores* de validação cruzada e do conjunto de teste computados. Em seguida, são apresentados os mapas geológicos preditivos e suas respectivas inconsistências, bem como mapas de probabilidades por classe. A quantificação da incerteza das previsões também é ilustrada por meio de mapas de entropia. Foi ainda realizada a interpretação do modelo selecionado a partir da análise do impacto de cada *feature* nas previsões. Por fim, são apresentados os resultados da descrição de fenômenos tipicamente associados a dados geoespaciais.

6.1 ANÁLISE EXPLORATÓRIA DOS DADOS

A Figura 23 apresenta as frequências absolutas das seis unidades litoestratigráficas aflorantes na área de estudo. O fato de as três classes mais frequentes (PP4esjc, PP4esb e PP3csbg) abrangerem mais de 86% da área e a grande discrepância entre as classes minoritária (MAcgg) e majoritária (PP4esjc) evidenciam uma situação de desbalanceamento de classes.

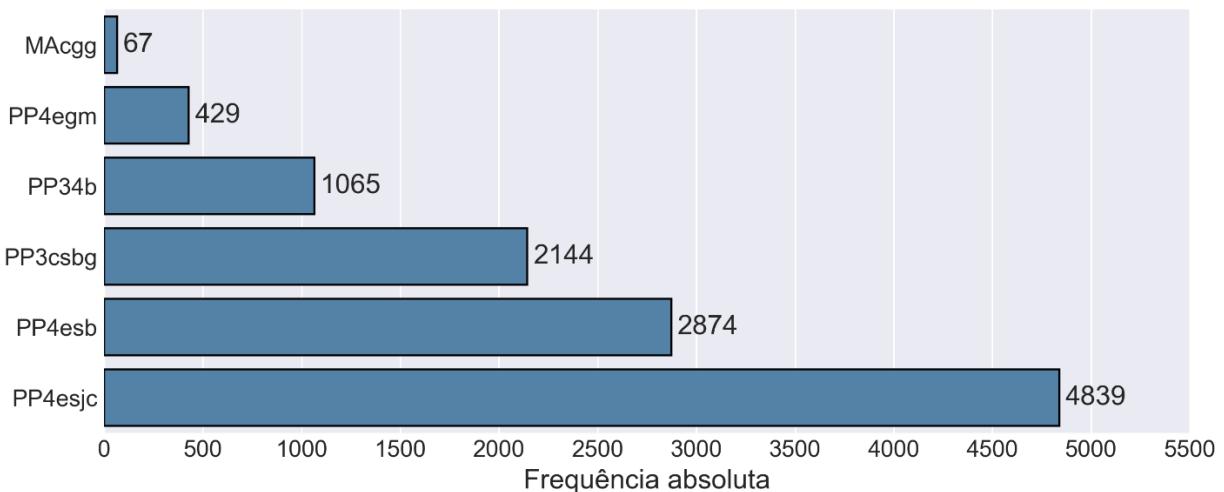


Figura 23 - Distribuição das seis unidades litoestratigráficas na área de estudo.

A Figura 24 mostra a distribuição das variáveis preditoras por unidade. Nota-se que as variáveis encontram-se em escalas (*i.e.* intervalo dos valores assumidos pelas variáveis) distintas. Os canais radiométricos, CT e o MDT apresentam maiores variações em suas

distribuições quando agrupados pelas unidades litoestratigráficas. Por outro lado, as bandas Landsat 8 evidenciam um padrão oposto, já que, quando individualizadas pelas classes, suas distribuições não variam significativamente.

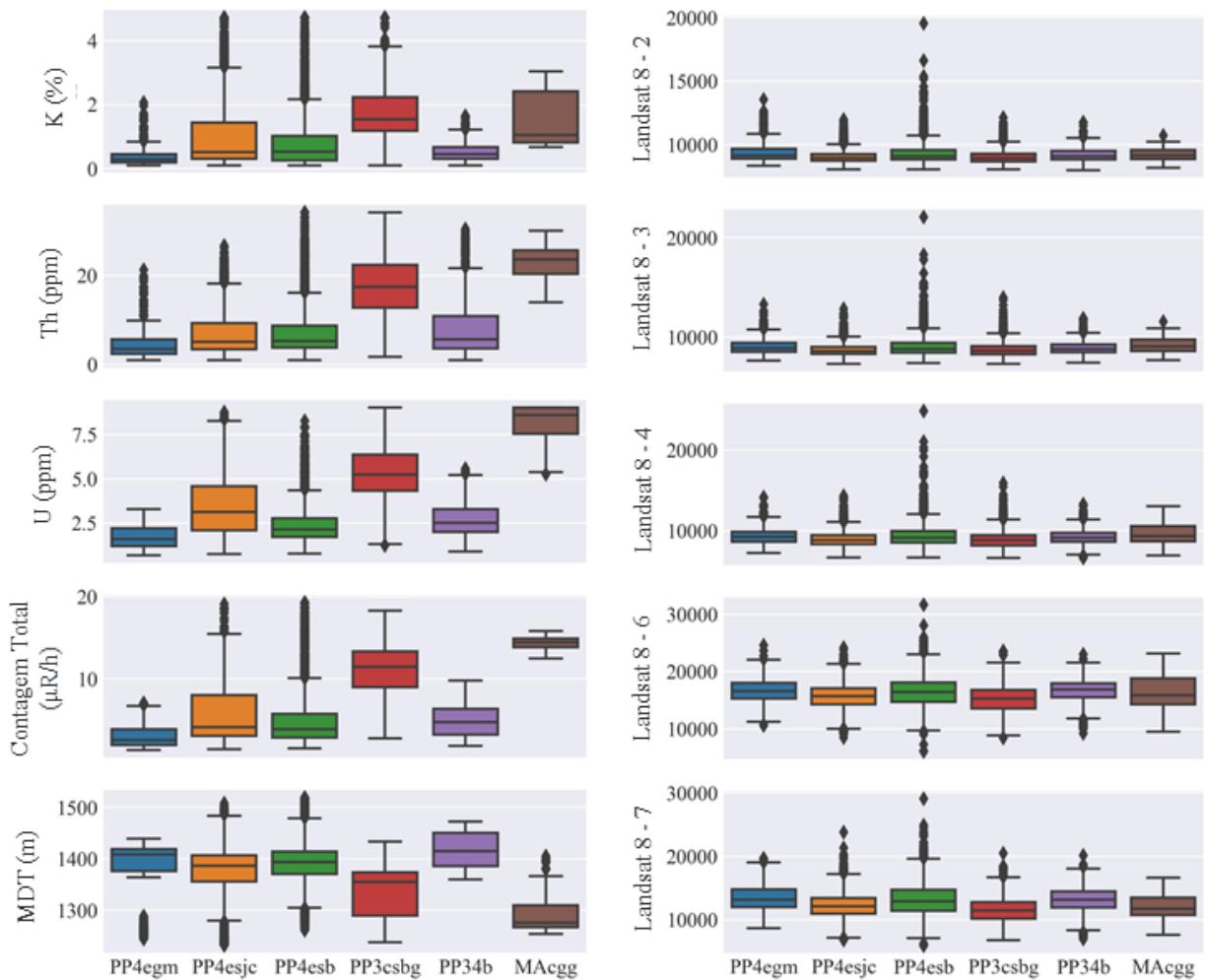


Figura 24 - Distribuição de algumas variáveis independentes por unidades litoestratigráficas.

A Figura 25 apresenta a matriz de correlação linear das variáveis independentes e evidencia uma forte correlação linear positiva entre os canais radiométricos e CT. O mesmo padrão é observado nas bandas Landsat 8 que, por sua vez, apresentam coeficientes de correlação de Pearson (r) iguais ou superiores a 0,8.

O Complexo Granito-Gnáissico (MACgg) mostra as maiores médias de U (8,049 ppm) e Th (22,976 ppm) e a segunda maior média de K (1,530 %). A distribuição de U nessa unidade é assimétrica negativa (Figura 26A), com um coeficiente de assimetria igual a -1,193. Já o histograma da variável Th é aproximadamente simétrico (Figura 26B), enquanto a distribuição de K é levemente assimétrica positiva (Figura 26C).

A Formação Barão de Guaicuí (PP3csbg) apresenta a maior média de K (1,738 %) dentre todas as unidades litoestratigráficas da área de estudo e a segunda maior média de Th (17,696 ppm). A distribuição de K, nessa classe, é tipicamente assimétrica positiva (Figura 27A), enquanto o histograma da variável radiométrica Th ilustra um padrão aproximadamente simétrico (Figura 27B).

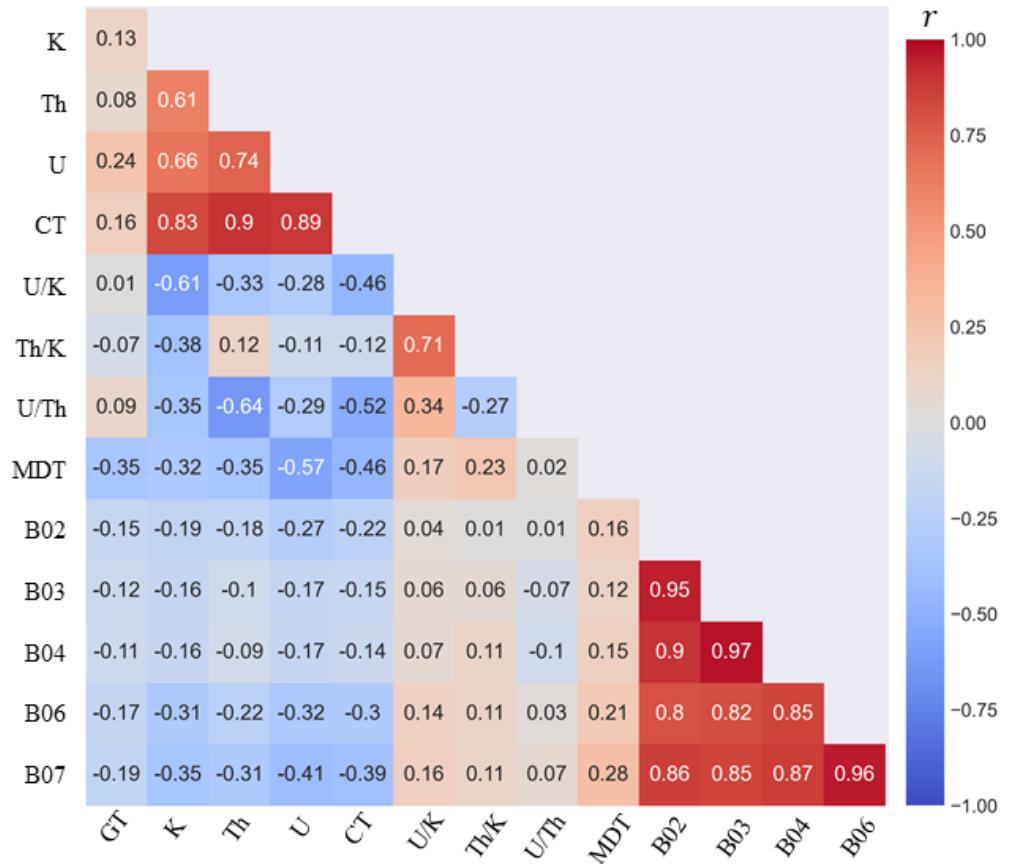


Figura 25 - Matriz de correlação linear entre as variáveis preditoras.

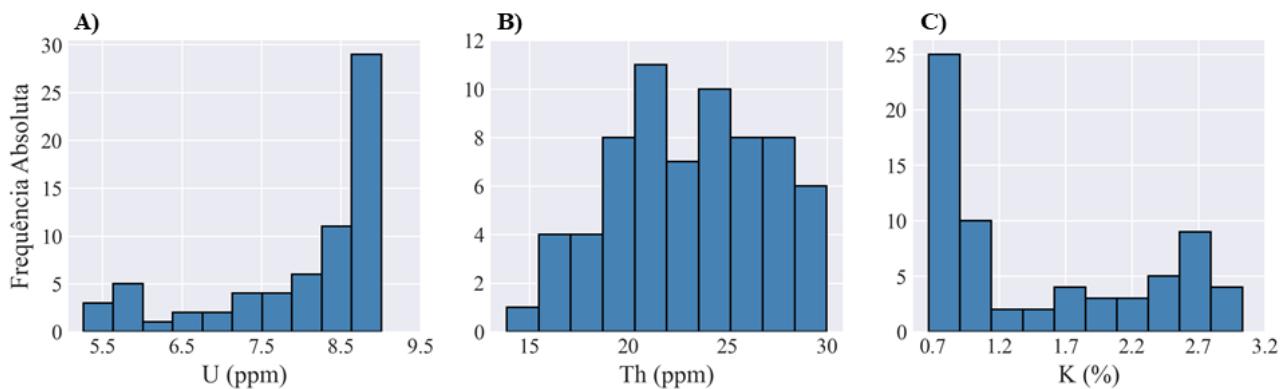


Figura 26 - Distribuições das variáveis (A) U (ppm), (B) Th (ppm) e (C) K (%) no Complexo Granito-Gnáissico.

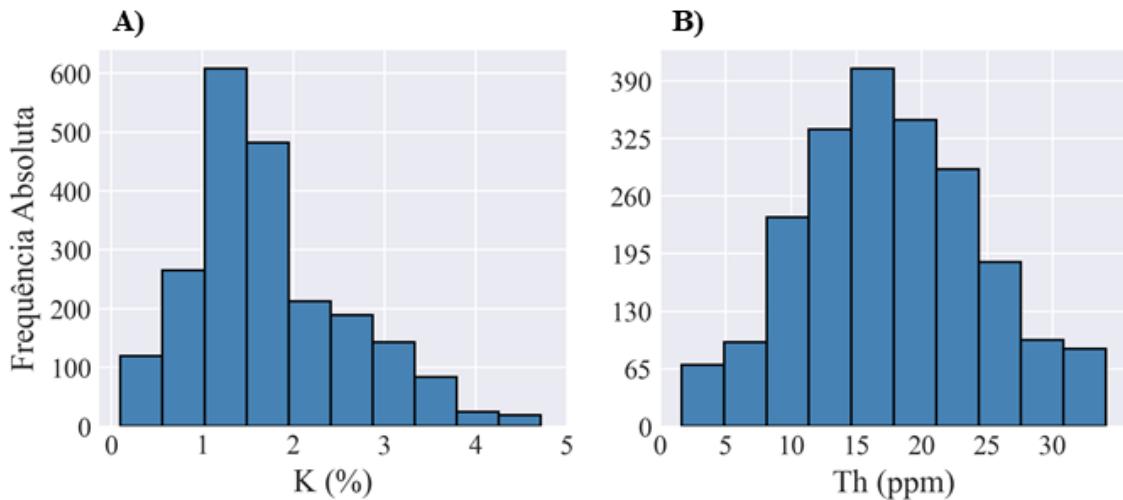


Figura 27 - Distribuições das variáveis (A) K (%) e (B) Th (ppm) na Formação Barão de Guaicuí.

A Formação Bandeirinha (PP34b) possui a maior elevação média dentre as unidades litoestratigráficas (1417,074 m) e 50% das amostras dessa classe mostram cotas superiores a 1414,320 m. Além disso, essa unidade apresenta a segunda maior média de razão U/K (5,74). Os histogramas de MDT e razão U/K encontram-se na Figura 28.

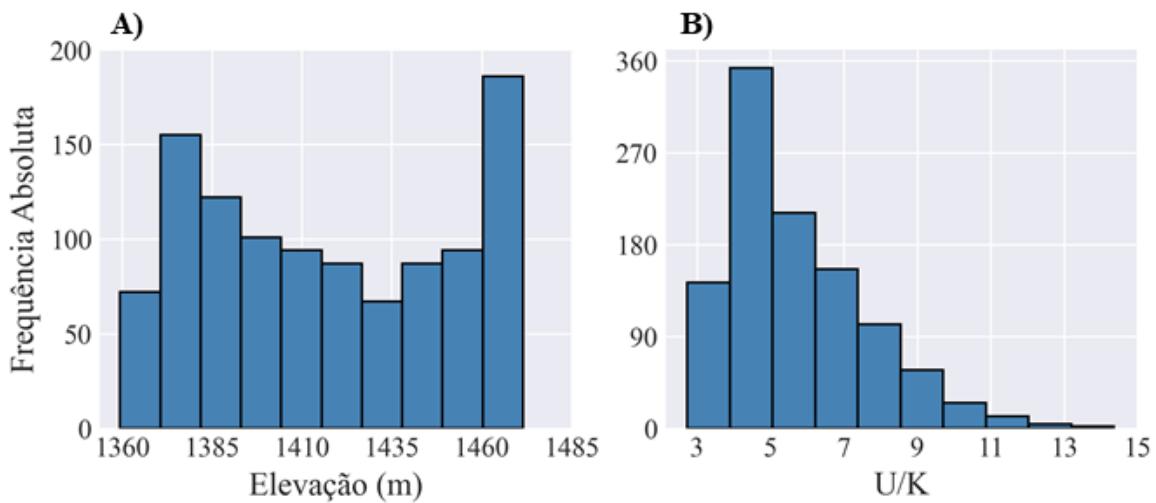


Figura 28 - Distribuições das variáveis (A) MDT (m) e (B) U/K na Formação Bandeirinha.

A Formação São João da Chapada (PP4esjc) mostra a maior média de razão U/Th (0,588) e a menor média de razão Th/K (9,986) dentre as unidades litoestratigráficas. Tanto a distribuição de U/Th (Figura 29A) quanto a de Th/K (Figura 29B) são assimétricas positivas com coeficientes de variação (C_v) superiores a 1,0.

A Formação Sopa-Brumadinho (PP4esb), por sua vez, possui a segunda menor média de U (2,435 ppm) dentre as unidades aflorantes na área. A distribuição de U (ppm) nessa

classe é fortemente assimétrica positiva (Figura 30), com um coeficiente de assimetria igual a 1,783.

Por fim, a Formação Galho do Miguel (PP4egm) apresenta as menores médias de U (1,713 ppm), K (0,381 %) e Th (4,528 ppm) dentre as unidades litoestratigráficas. As distribuições desses canais radiométricos nessa classe mostram-se assimétricas positivas (Figura 31). Além disso, o histograma da variável MDT apresenta um padrão bimodal, com um *gap* entre 1300 m e 1350 m, aproximadamente (Figura 31B).

Os sumários estatísticos geral e agrupado pelas unidades litoestratigráficas encontram-se nos anexos II e III, respectivamente.

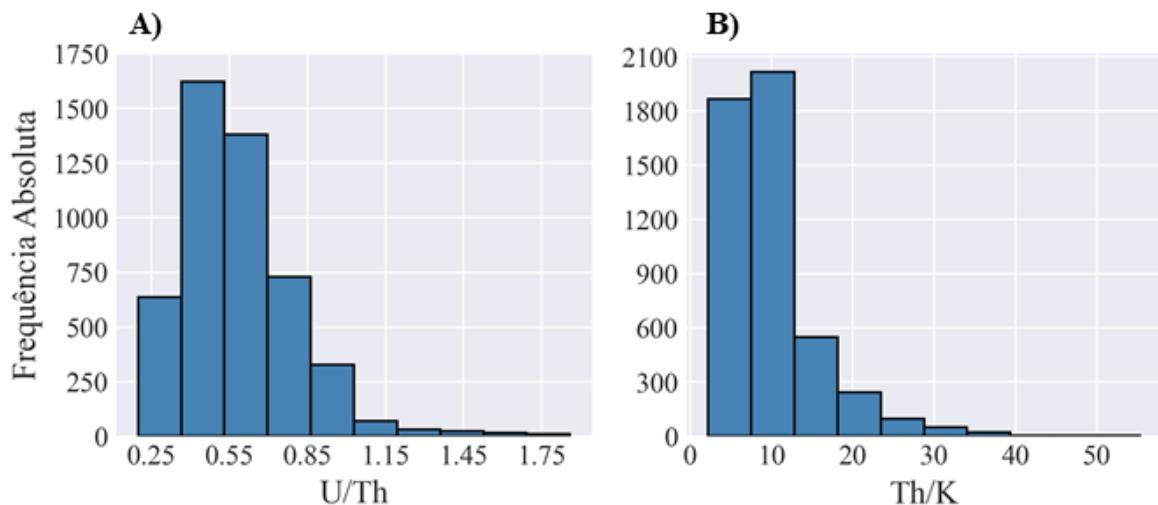


Figura 29 - Distribuições das variáveis (A) U/Th e (B) Th/K na Formação São João da Chapada.

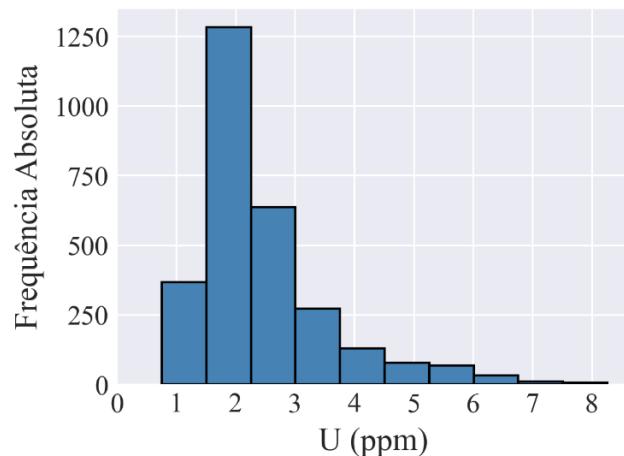


Figura 30 - Distribuição da variável U (ppm) na Formação Sopa-Brumadinho.

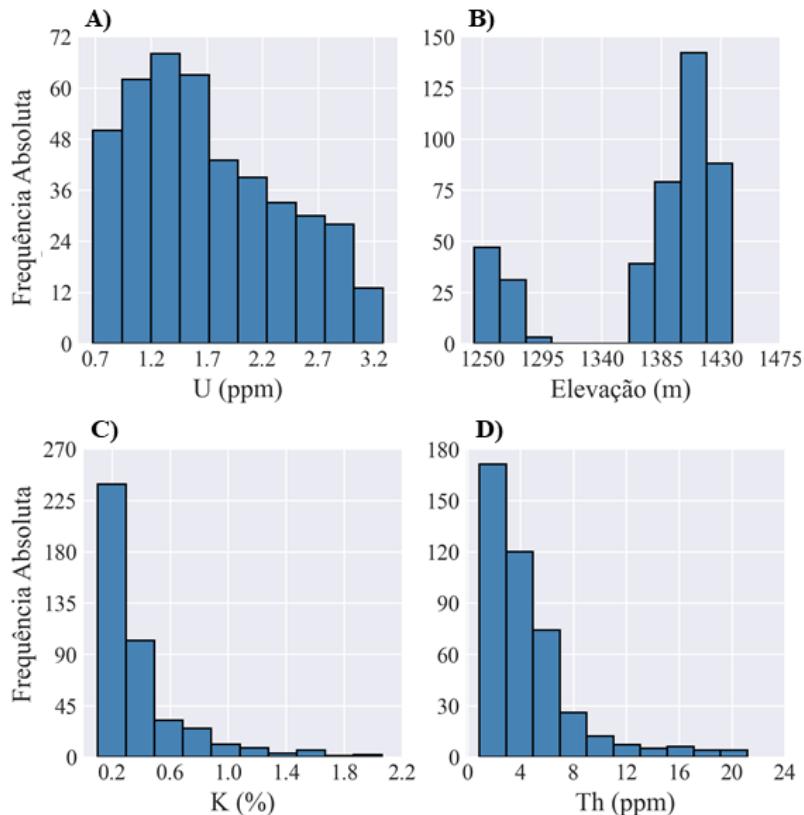


Figura 31 - Distribuições das variáveis (A) U (ppm), (B) MDT (m), (C) K (%) e (D) Th (ppm) na Formação Galho do Miguel.

6.2 SEPARAÇÃO ENTRE TREINO E TESTE

Após a separação, 7% das observações constituem o conjunto de treino (T_a), ao passo que 93% compõem o conjunto de teste (T_b) (Tabela 11).

Tabela 11 - Número de instâncias em T_a e T_b após a separação entre treino e teste.

	Instâncias pertencentes à T_a	Instâncias pertencentes à T_b
Frequência Absoluta	797	10621
Frequência Relativa	7%	93%

A Figura 32 apresenta as frequências absolutas das unidades litoestratigráficas no conjunto de treino. Ainda que MACGG permaneça como unidade minoritária (47 instâncias), nota-se que o desbalanceamento de classes foi mitigado. Isso é evidenciado através de uma comparação entre as distribuições das unidades no banco de dados (Figura 23) e no conjunto de treino (Figura 32).

A Figura 33 ilustra a distribuição espacial das observações presentes no conjunto de treino. Percebe-se que os exemplos de treino, embora representem apenas 7% do banco de dados, abrangem toda a área de mapeamento.

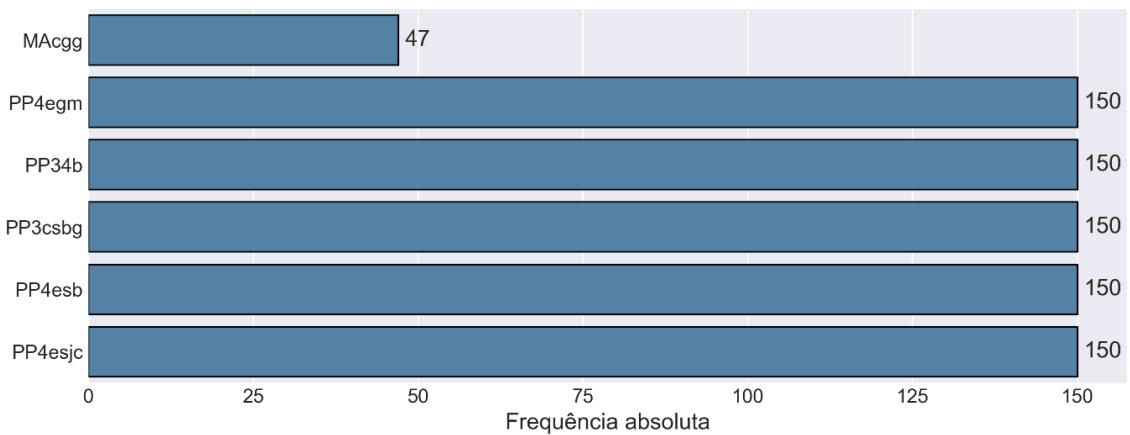


Figura 32 – Distribuição das unidades litoestratigráficas no conjunto de treino após a separação dos dados.

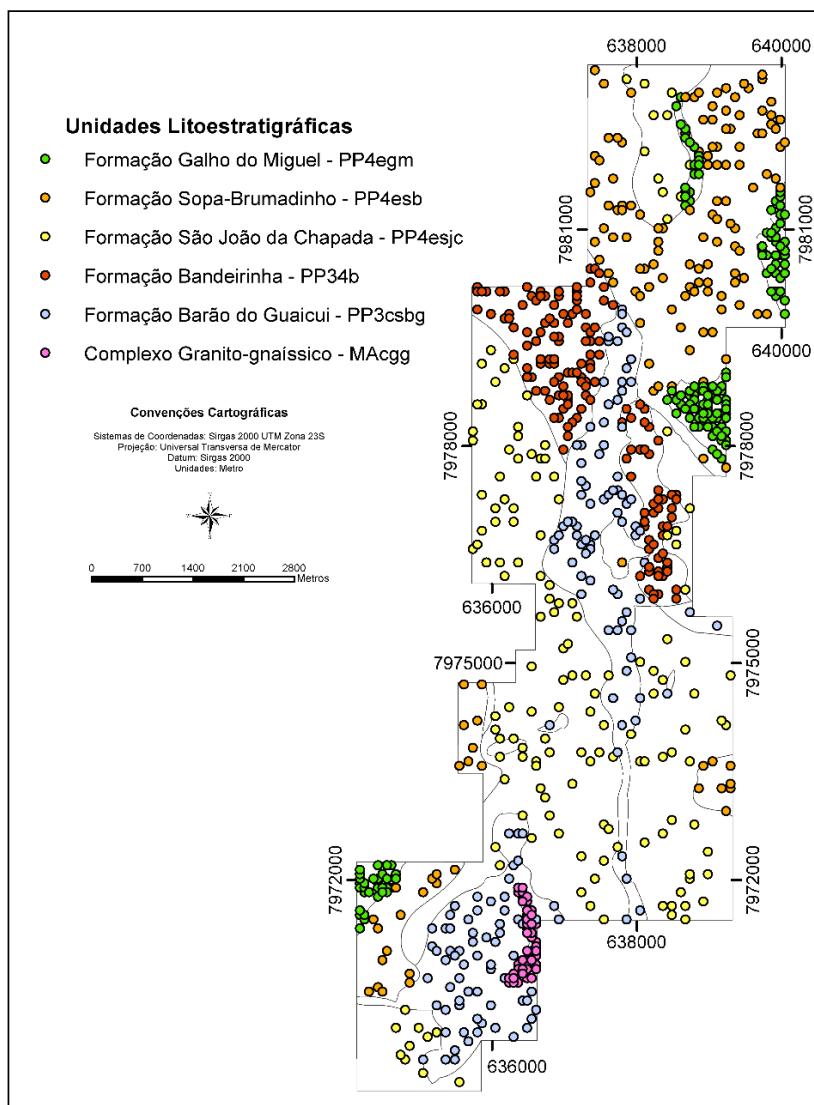


Figura 33 - Distribuição espacial das amostras de treino. As cores representam as unidades litoestratigráficas.

6.3 PRÉ-PROCESSAMENTO

O pré-processamento dos dados de treino foi dividido em três etapas sequenciais, sendo elas o escalonamento das *features*, a redução da dimensionalidade e a superamostragem. Essas etapas foram então encadeadas em um *pipeline*, visando automatizar a rotina de pré-processamento para cada um dos algoritmos.

Conforme observado na Figura 24, as variáveis independentes possuem escalas distintas e, por isso, foram estandardizadas. A Tabela 12 apresenta o sumário estatístico das variáveis independentes escalonadas no conjunto de treino. Nota-se que, após a estandardização, todas as *features* seguem uma distribuição normal padrão, com média e desvio padrão iguais a 0 e 1, respectivamente.

Tabela 12 - Sumário estatístico das variáveis independentes estandardizadas no conjunto de treino.

Estatísticas	GT	K	Th	U	CT	U/K	Th/K	U/Th	MDT	B02	B03	B04	B06	B07
Contagem	797	797	797	797	797	797	797	797	797	797	797	797	797	797
X̄	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Min	-0,97	-0,97	-1,12	-1,31	-1,2	-1,32	-0,97	-1,8	-2,53	-1,75	-1,77	-2,15	-4,02	-2,91
P10%	-0,75	-0,88	-0,93	-1	-0,99	-0,95	-0,71	-1,07	-1,8	-0,97	-1,02	-1,13	-1,28	-1,31
P25%	-0,62	-0,73	-0,78	-0,74	-0,8	-0,68	-0,54	-0,68	-0,44	-0,61	-0,65	-0,61	-0,61	-0,62
P50%	-0,33	-0,35	-0,39	-0,31	-0,37	-0,24	-0,29	-0,18	0,2	-0,24	-0,21	-0,1	0,03	-0,06
P75%	0,22	0,39	0,66	0,6	0,65	0,35	0,15	0,49	0,72	0,39	0,42	0,44	0,64	0,66
P99,5%	5,07	3,85	3,17	2,74	2,38	3,72	5,43	3,28	1,7	4,56	4,13	3,45	2,44	2,81
Max	8,19	4,54	3,17	2,74	2,79	5,27	7,34	6,23	2,56	6,34	6,45	6,45	2,9	3,07

\bar{X} = média; S = desvio padrão; Min = mínimo; P10% = percentil 10; P25% = percentil 25; P50% = mediana; P75% = percentil 75; P99,5% = percentil 99,5; Max = máximo.

Durante a etapa seguinte, as bandas Landsat 8 foram submetidas à técnica de ACP, visando a redução do número de dimensões e da forte correlação linear positiva entre essas variáveis (Figura 25). A Figura 34 ilustra uma matriz de correlação linear entre as bandas Landsat 8 estandardizadas e as cinco componentes principais obtidas. Observa-se que a primeira componente principal (PC1) apresenta uma forte correlação linear positiva ($r > 0,9$) com todas as cinco bandas escalonadas. Por outro lado, as demais componentes principais mostram correlações lineares (positivas ou negativas) fracas com essas bandas.

A Figura 35 representa o percentual de variância explicada por cada uma das cinco componentes principais. As barras mostram a contribuição individual de cada componente na explicação da variabilidade total das cinco bandas Landsat 8, enquanto a linha preta representa a variância explicada acumulada pelas componentes. Nota-se que apenas a PC1 é capaz de explicar mais de 90% da variabilidade contida nas bandas Landsat 8. As demais

componentes somadas, por outro lado, pouco contribuem para a explicação dessa variância total (< 10%).

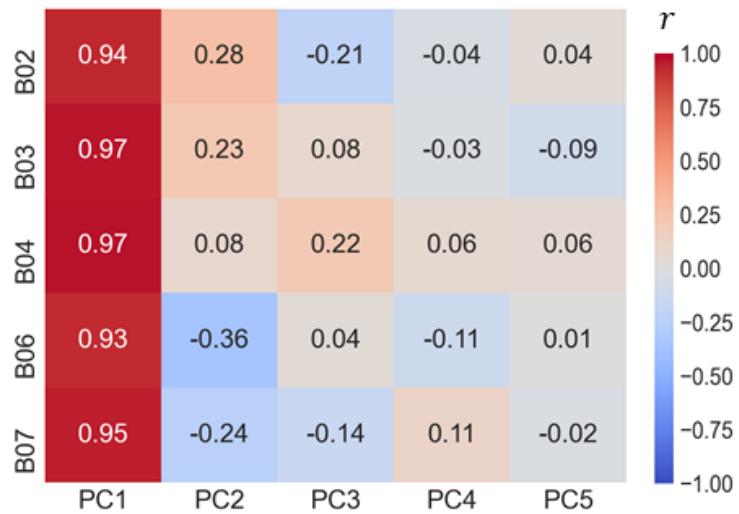


Figura 34 - Matriz de correlação linear entre as bandas Landsat 8 estandardizadas e as componentes principais.

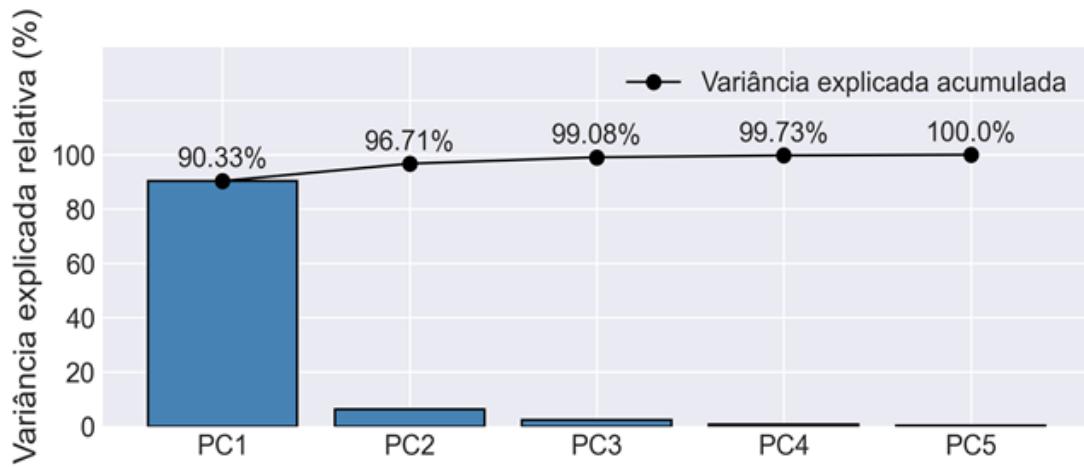


Figura 35 - Variância explicada relativa de cada uma das cinco componentes principais. A linha preta representa a variância explicada acumulada.

Com base nos resultados obtidos (Figura 34 e Figura 35), optou-se por utilizar apenas a PC1, de modo que o número de dimensões foi reduzido de 14 para 10. A Figura 36 mostra a matriz de correlação linear entre as variáveis independentes após a etapa de redução da dimensionalidade. Nota-se que houve uma redução no número de variáveis preditoras fortemente correlacionadas entre si quando comparado com a matriz de correlação linear pré-ACP (Figura 25). Entretanto, como CT ainda mostra uma forte correlação linear positiva com os canais radiométricos (U, Th e K), optou-se por removê-la.

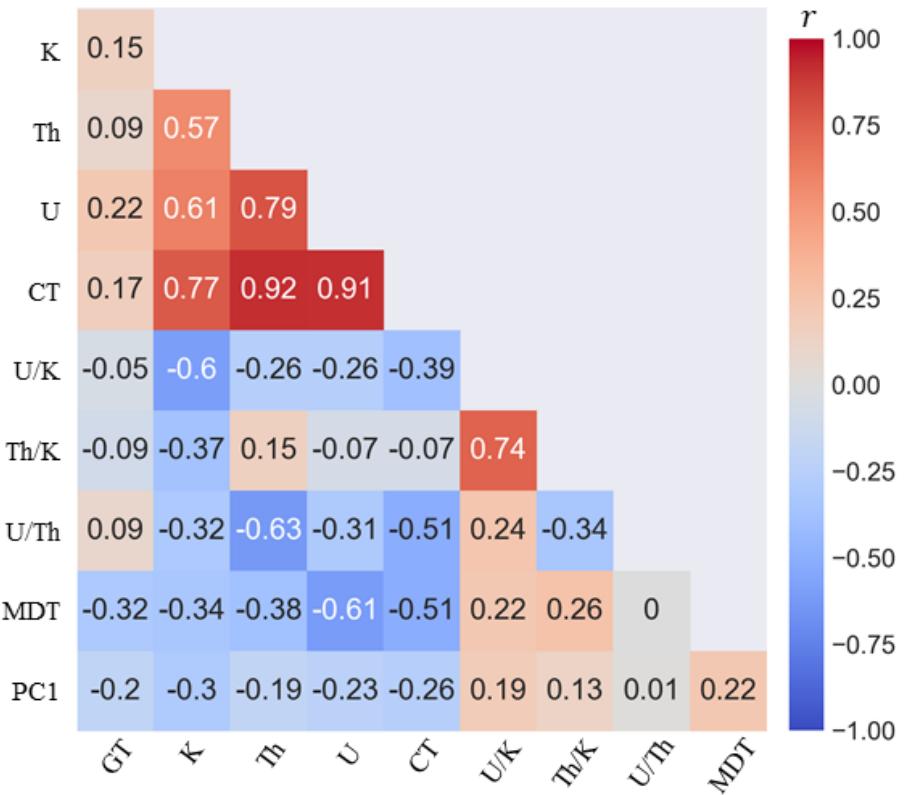


Figura 36 - Matriz de correlação linear entre as variáveis independentes após a redução da dimensionalidade.

Durante a última etapa de pré-processamento, aplicou-se o SMOTE, visando a superamostragem da classe minoritária no conjunto de treino. A Figura 37 apresenta as frequências absolutas de cada uma das unidades litoestratigráficas após a geração de 103 amostras sintéticas da unidade MACgg. Nota-se que, ao final desta etapa, as classes do conjunto de treino se encontram perfeitamente balanceadas.

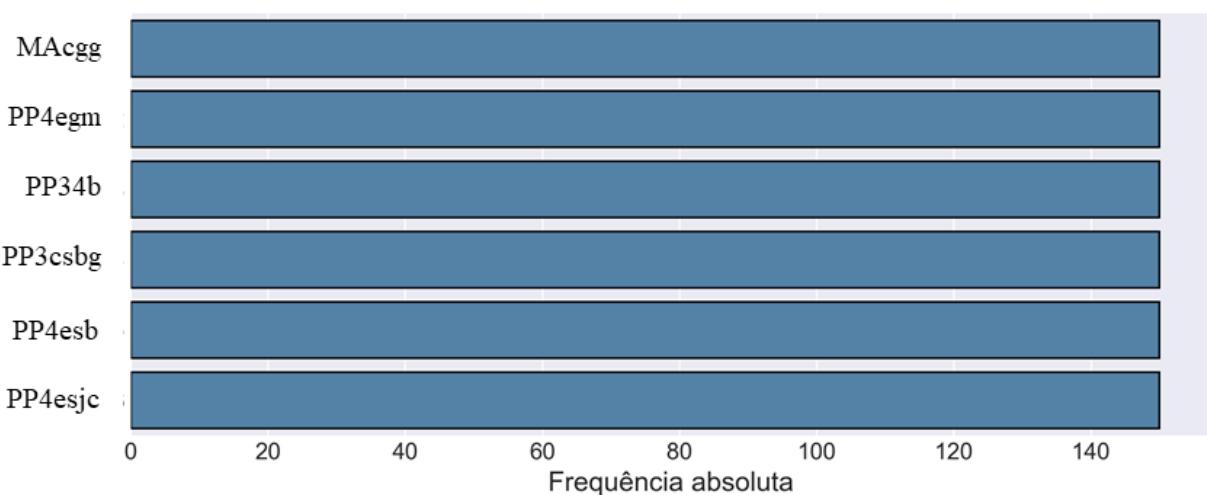


Figura 37 - Distribuição das unidades litoestratigráficas no conjunto de treino, após a superamostragem da unidade MACgg.

6.4 OTIMIZAÇÃO DOS HIPERPARÂMETROS

Os hiperparâmetros dos algoritmos RF, XGB e MLP foram otimizados por meio da técnica *Random Search*, ao passo que, no caso dos demais classificadores, aplicou-se o método *Grid Search*. Em ambos os casos, a métrica de referência adotada foi *F1-score*. A Tabela 13 mostra os hiperparâmetros ótimos selecionados para cada um dos algoritmos, bem como seus respectivos valores de *F1-score*. Nota-se que RF, XGB e MLP apresentam os maiores valores de *F1-score*, enquanto os menores *scores* estão associados aos algoritmos NB e RL.

Tabela 13 - Hiperparâmetros ótimos e valores de F1-score para cada algoritmo.

Algoritmos	Hiperparâmetros	Valores Ótimos	F1-score	Algoritmos	Hiperparâmetros	Valores Ótimos	F1-score	
RL	<i>solver</i>	newton-cg	0,59	KNN	<i>n_neighbors</i>	3		
	<i>C</i>	2,15E-02			<i>weights</i>	distance	0,68	
SVM	<i>C</i>	1,00E+3	0,71		<i>p</i>	1		
	<i>gamma</i>	auto			<i>n_estimators</i>	500		
NB	<i>kernel</i>	rbf	0,51	RF	<i>max_depth</i>	25		
	<i>var_smoothing</i>	1E-10			<i>criterion</i>	entropy	0,78	
DT	<i>criterion</i>	entropy	0,68	XGB	<i>min_samples_split</i>	5		
	<i>max_depth</i>	25			<i>min_samples_leaf</i>	1		
	<i>min_samples_split</i>	2			<i>eta</i>	0,025		
	<i>min_samples_leaf</i>	1			<i>learning_rate</i>	0,3		
MLP	<i>hidden_layer_sizes</i>	(20,20)	0,76	XGB	<i>gamma</i>	0,05		
	<i>activation</i>	tanh			<i>max_depth</i>	25		
	<i>solver</i>	adam			<i>subsample</i>	0,9	0,77	
	<i>alpha</i>	0,1			<i>colsample_bytree</i>	1		
	<i>learning_rate</i>	adaptive			<i>min_child_weight</i>	5		
	<i>learning_rate_init</i>	0,01			<i>reg_lambda</i>	0,001		
	<i>max_iter</i>	200			<i>alpha</i>	0,1		

6.5 PERFORMANCE DOS CLASSIFICADORES

Nesta seção, é apresentada uma comparação entre as estimativas de performance obtidas na VCKF com cinco *folds* e as performances dos modelos no conjunto de teste.

A Figura 38 exibe três mapas de calor. Os mapas das porções superior e central apresentam, respectivamente, os *scores* da VCKF e do conjunto de teste para cada um dos oito classificadores avaliados, considerando as métricas *F1-score*, precisão, *recall* e acurácia. Já o mapa de calor da porção inferior contém as diferenças entre os *scores* da VCKF e do conjunto de teste. Valores negativos e positivos indicam estimativas de performance pessimistas e otimistas, respectivamente.

Nota-se, a partir da observação dos *scores* de VCKF e teste, que os classificadores mais flexíveis, como RF, XGB e MLP apresentaram as melhores performances, enquanto modelos mais simples, como RL e NB mostraram performances inferiores. Já DT, KNN e SVM, por sua vez, exibiram performances intermediárias. Observa-se ainda que as diferenças entre as estimativas e os *scores* de teste foram sempre inferiores a 0,10. De forma geral, as estimativas fornecidas pela VCKF tendem a ser ligeiramente otimistas, com exceção das estimativas de precisão para os modelos RL e NB, que foram relativamente pessimistas.

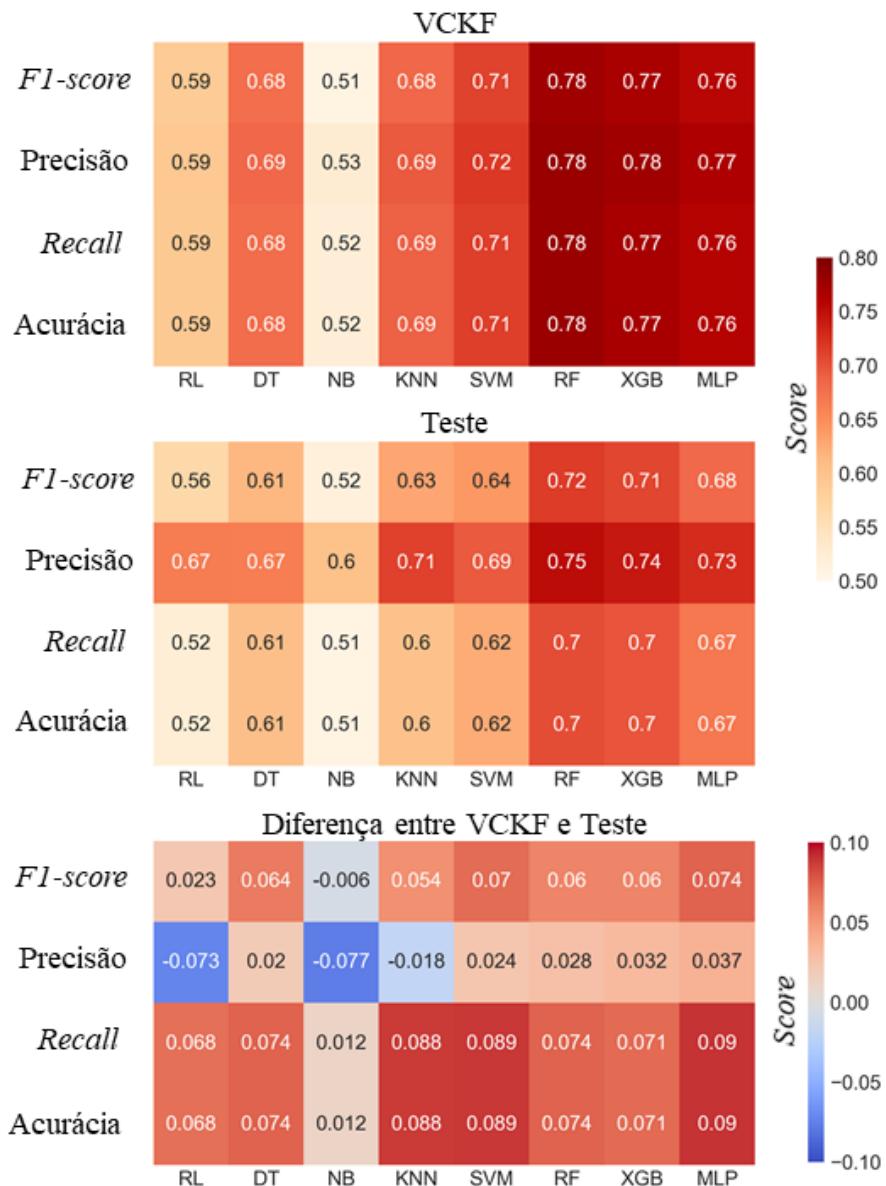


Figura 38 - Scores dos classificadores obtidos na VCKF (superior) e no conjunto de teste (centro). O mapa de calor inferior apresenta as diferenças entre os scores de VCKF e teste.

A Figura 39 elenca os modelos em ordem decrescente de acordo com os valores de *F1-score* obtidos na VCKF (superior) e no conjunto de teste (inferior). Em geral, ainda que as estimativas de *F1-score* sejam ligeiramente otimistas, como já evidenciado na Figura 38, é

possível perceber que os classificadores foram elencados na mesma ordem em ambos os casos.



Figura 39 - Classificadores elencados em ordem decrescente de acordo com os valores de F1-score obtidos na VCKF (superior) e no conjunto de teste (inferior).

A Figura 40 compara as matrizes de confusão do conjunto de teste associadas a cada um dos oito modelos. Cada linha representa o percentual de instâncias pertencentes a uma unidade do mapa integrado de campo, ao passo que cada coluna mostra o percentual de previsões referentes a uma unidade do mapa preditivo. Nesse sentido, a análise dessas matrizes permite avaliar a performance de cada modelo por unidade litoestratigráfica.

Nota-se que os modelos obtiveram melhores performances na classificação das unidades menos frequentes (*i.e.* menor número de instâncias), como MACgg, com 67 amostras, e PP4egm, com 429 observações. Essas unidades foram frequentemente confundidas com PP3csbg e PP4esb, respectivamente. Por outro lado, os classificadores apresentaram maiores dificuldades na predição das classes mais aflorantes (*i.e.* maior número de instâncias), como é o caso das unidades PP4esjc, com 4839 amostras, e PP4esb, com 2874 exemplos. Ambas as unidades são comumente confundidas entre si, de modo que PP4esjc é, também, classificada incorretamente como PP3csbg ou PP34b.

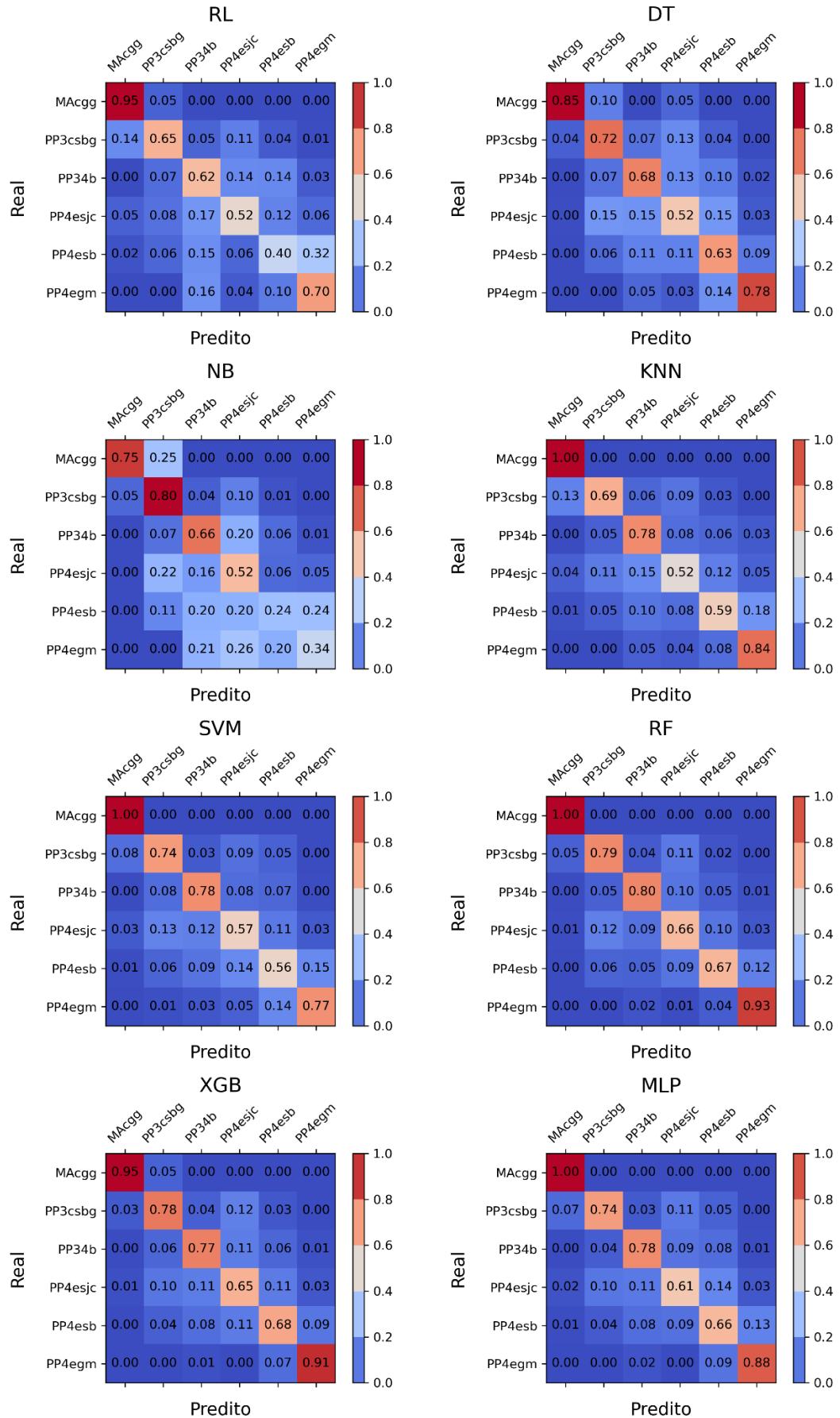


Figura 40 - Matrizes de confusão do conjunto de teste associadas a cada modelo.

6.6 MAPAS GEOLÓGICOS PREDITIVOS

A Figura 41 exibe os oito mapas geológicos preditivos gerados e seus respectivos erros de classificação. Essas inconsistências ocorrem quando, para uma determinada instância, a classe predita é diferente da unidade litoestratigráfica interpretada no mapa integrado de campo (Figura 11). Nesse sentido, elas podem estar associadas a erros de predição, erros de interpretação em campo ou ambos e, portanto, não necessariamente implicam em inconsistências do modelo.

Nota-se que, em geral, todos os modelos cometem menos erros de classificação na porção centro-oeste da área de mapeamento. Ainda que os mapas preditivos RL e NB apresentem mais inconsistências do que os demais, percebe-se que as previsões geradas por esses modelos são espacialmente mais contínuas. Por outro lado, os mapas preditivos gerados pelos modelos SVM, DT, KNN e MLP apresentam-se mais ruidosos (*i.e.* menos contínuos), embora possuam menos inconsistências do que RL e NB. Os modelos RF e XGB, por sua vez, geram mapas com maior continuidade espacial do que SVM, DT, KNN e MLP e cometem menos erros de classificação do que os demais.

A Figura 42 e a Figura 43 ilustram os mapas de probabilidade por unidade litoestratigráfica relativos aos modelos RF e XGB, respectivamente. Nesse sentido, quanto maior é a probabilidade de uma determinada classe, maior é a chance de o algoritmo predizê-la. Nota-se que, para cada uma das unidades, os dois modelos exibem padrões espaciais de probabilidade semelhantes. Entretanto, as probabilidades preditas pelo XGB, para cada uma das classes, são mais altas (*i.e.* cores mais quentes) do que aquelas preditas pelo RF.

As maiores probabilidades da unidade MACgg se localizam exclusivamente na porção sul da área, ao passo que, na classe PP3csbg, os valores mais altos se situam nas porções central e sul. A unidade PP34b, por sua vez, exibe probabilidades mais altas nas regiões noroeste e centro-leste, enquanto PP4esjc apresenta seus valores mais elevados nas porções oeste, centro-sul e norte. No caso da classe PP4esb, as probabilidades mais elevadas situam-se à norte, nordeste e sudoeste da área. Por fim, a unidade PP4egm exibe os valores mais altos de probabilidade nas regiões nordeste e, localmente, sudoeste.

O Anexo IV apresenta o mapa preditivo resultante do modelo XGB pós-processado com a aplicação de filtros de redução de ruídos. Nesse mesmo anexo são expostos mapas de probabilidade por unidades e um mapa de entropia, sendo todos eles também associados ao

modelo XGB. São mostrados ainda os mapas de alguns sensores remotos, bem como o mapa geológico integrado.

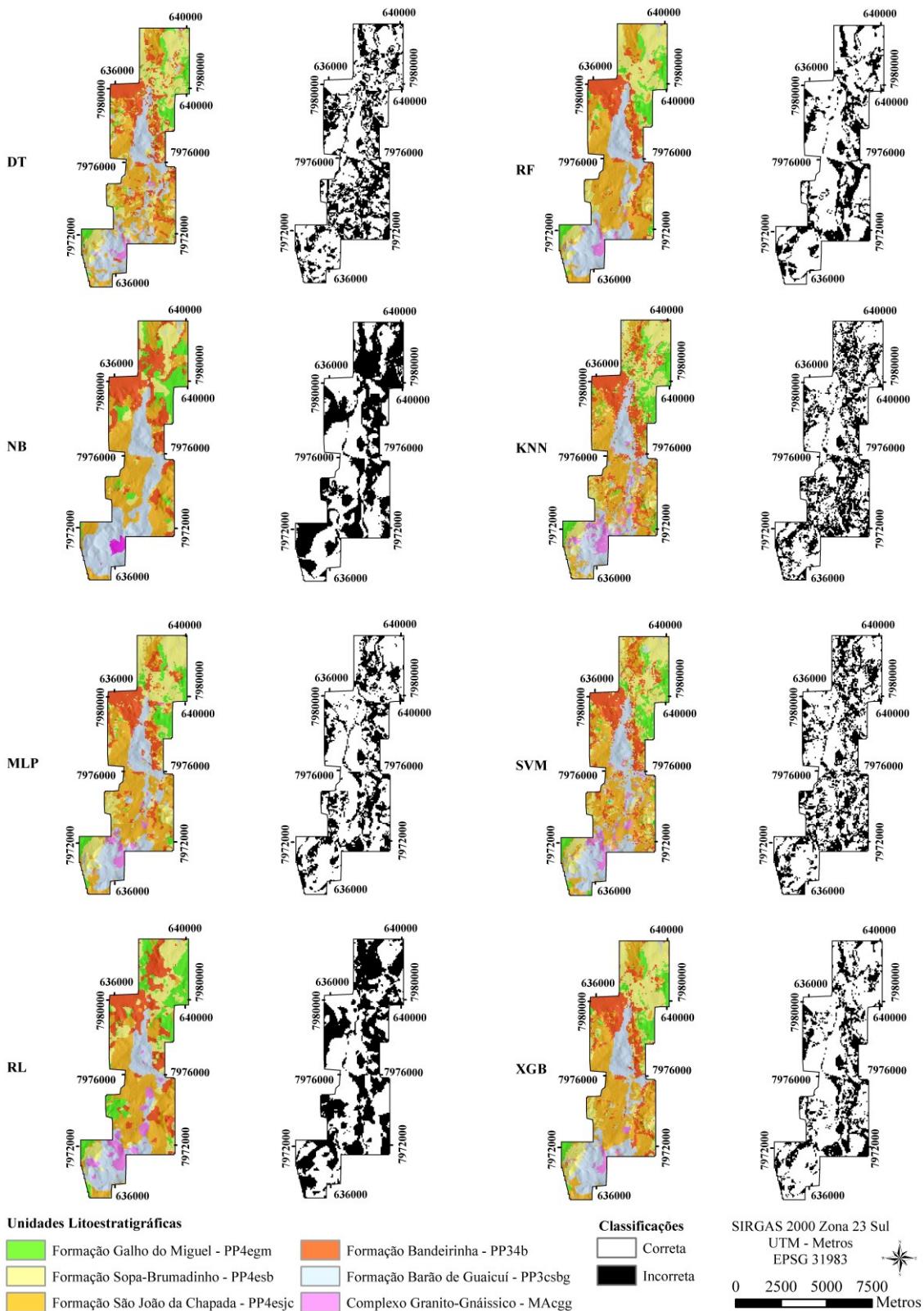


Figura 41 - Mapas geológicos preditivos e seus respectivos erros de classificação.

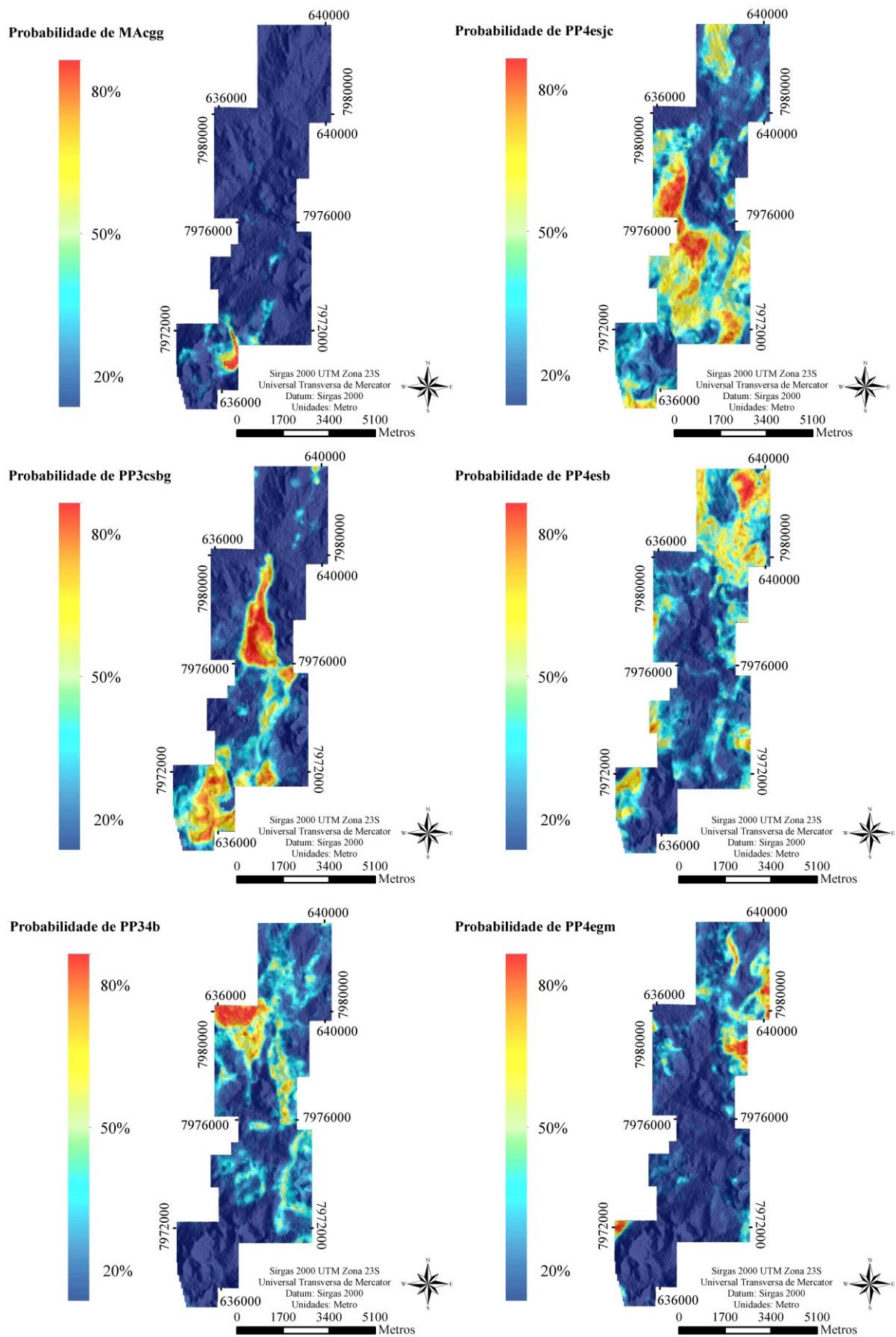


Figura 42 - Mapas de probabilidade por classe gerados pelo modelo RF.

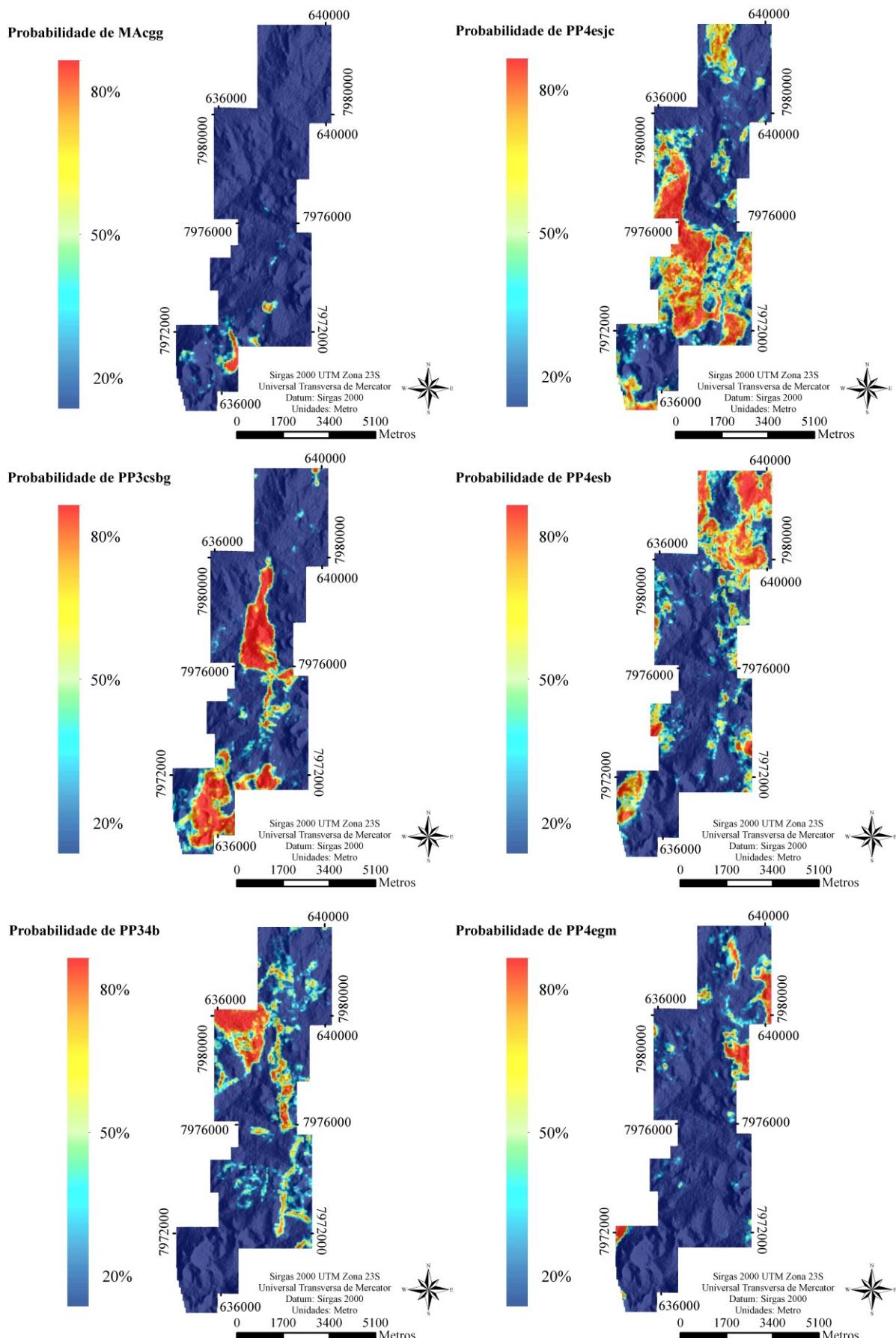


Figura 43 - Mapas de probabilidade por classe gerados pelo modelo XGB.

6.7 QUANTIFICAÇÃO DA INCERTEZA DAS PREDIÇÕES

A Figura 44 compara as incertezas das previsões dos modelos *Random Forests* e *XGBoost* com base na entropia da informação. Nota-se que os maiores valores de entropia, em ambos os casos, se associam a regiões geológicas mais complexas e mostram uma relação espacial com os contatos geológicos. Além disso, o mapa de entropia do modelo *Random Forests* exibe, em geral, valores de entropia maiores do que aqueles associados ao modelo *XGBoost*.

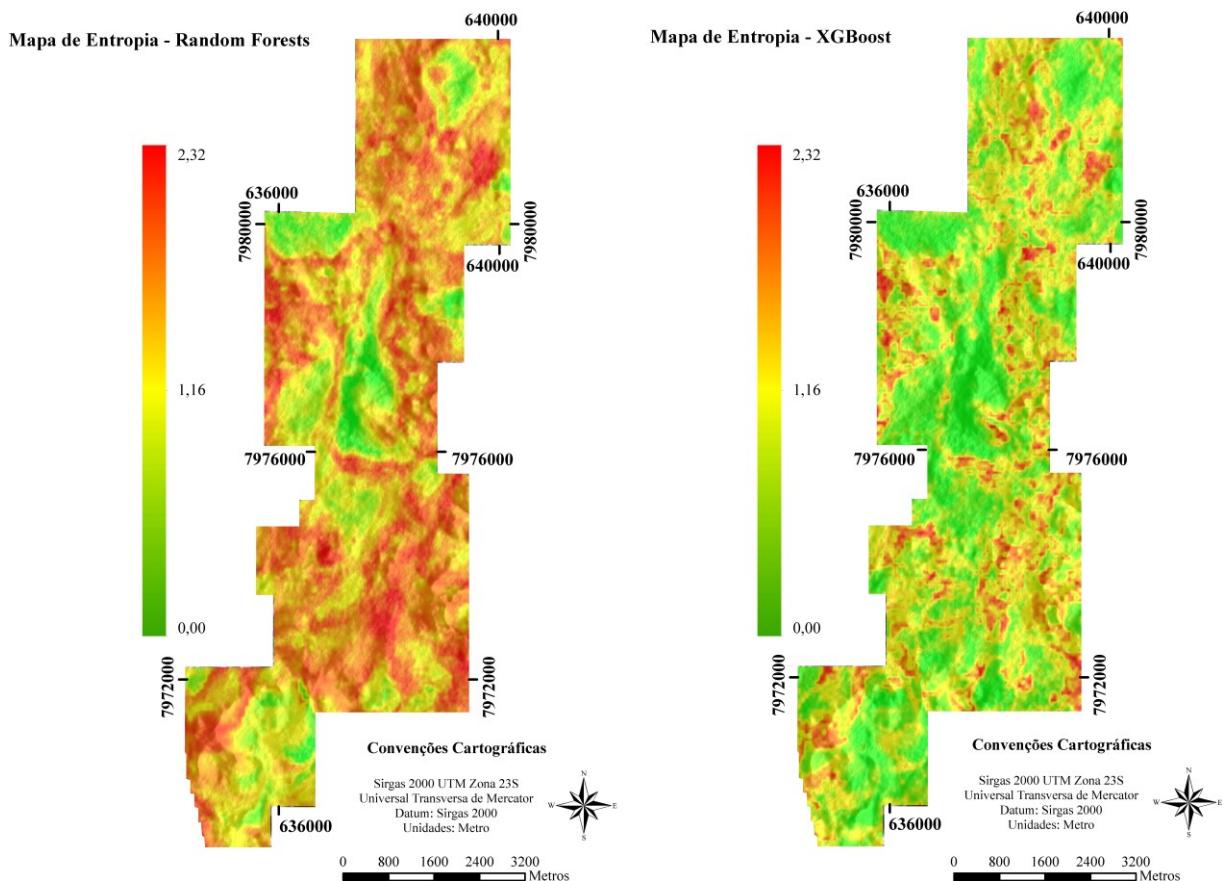


Figura 44 - Mapas de entropia dos modelos Random Forests e XGBoost.

6.8 INTERPRETAÇÃO DO MODELO XGBOOST

A Figura 45 descreve a importância das variáveis independentes nas previsões das unidades geradas pelo modelo *XGBoost*. Percebe-se que, no geral, as *features* mais importantes são os canais radiométricos (*i.e.* U, K e Th) e a razão U/Th, enquanto GT e PC1 exercem os menores impactos nas classificações.

Nota-se que as variáveis independentes de maior impacto nas previsões da unidade MAcgg foram U e Th, ao passo que, no caso de PP3csbg, K foi a *feature* mais importante. Já as previsões das classes PP34b e PP4esjc foram mais influenciadas pela elevação (MDT) e

pela razão U/Th, respectivamente. Além disso, nota-se que U é a variável de maior impacto nas previsões das unidades PP4esb e PP4egm.

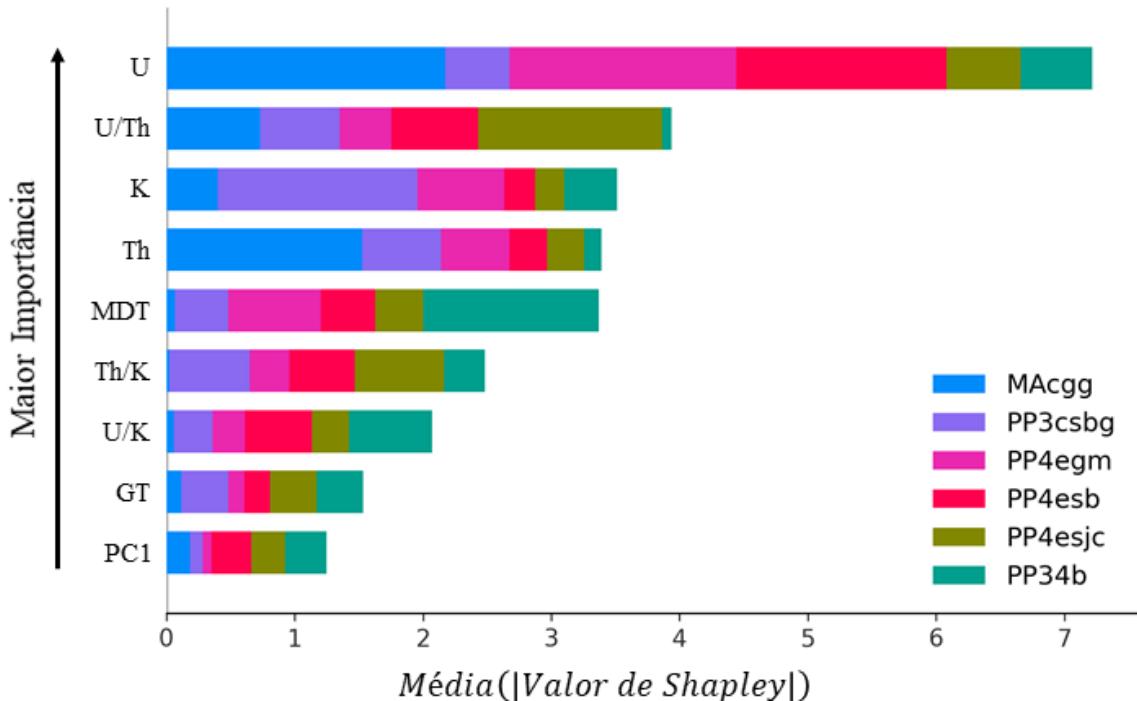


Figura 45 - Impacto das variáveis independentes nas previsões do modelo XGBoost.

Os *summary plots* por unidade litoestratigráfica apresentados a seguir relacionam os valores das *features* (altos ou baixos) aos impactos (positivos ou negativos) nas previsões realizadas pelo modelo *XGBoost*.

A Figura 46 sugere que valores altos de U e Th impactam positivamente na previsão da unidade Complexo Granito-Gnáissico (MAcgg). Nota-se ainda que as razões U/K e Th/K pouco contribuem para a classificação dessa unidade.

A Figura 47 indica que K é a variável de maior importância na previsão da Formação Barão de Guaicuí (PP3csbg), fato também observado na Figura 45. Nesse sentido, valores elevados de K favorecem a previsão dessa unidade, enquanto valores baixos dessa variável radiométrica levam o modelo a não predizê-la. Outro ponto em destaque relaciona-se à elevação (MDT) de modo que, mesmo não sendo uma das *feature* mais importantes, seus altos valores desfavorecem a previsão dessa unidade.

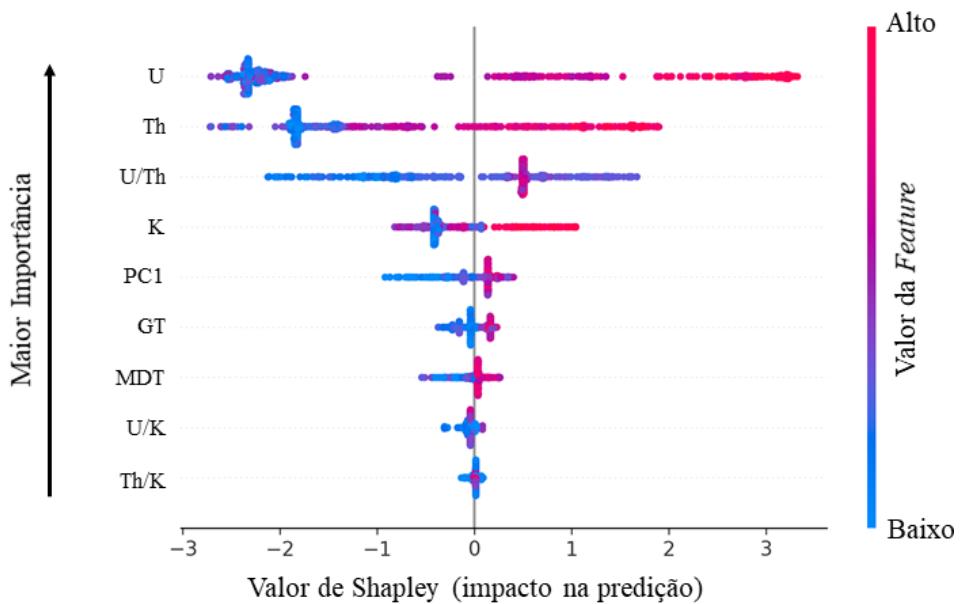


Figura 46 - Summary plot do Complexo Granito-Gnássico.

A Figura 48 mostra que altos valores de elevação (MDT) impactam positivamente na predição da Formação Bandeirinha (PP34b), enquanto valores baixos dessa variável levam o modelo a não predizer essa unidade. Nota-se que Th e U/Th são as variáveis que menos influenciam na predição dessa classe.

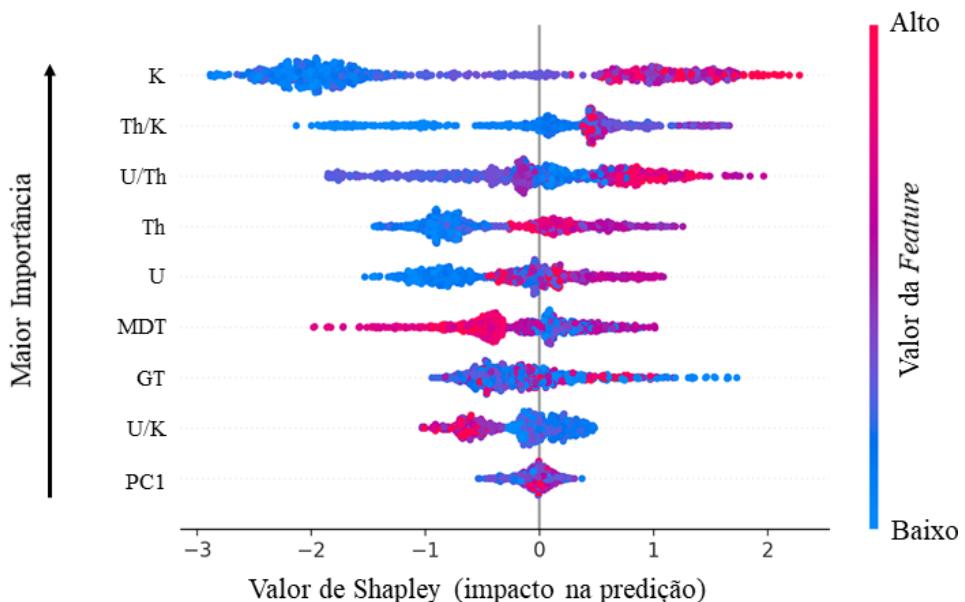


Figura 47 - Summary plot da Formação Barão de Guaicuí.

A Figura 49 sugere que elevados valores de U/Th, elevação (MDT) e U favorecem a predição da Formação São João da Chapada (PP4esjc), sendo a razão radiométrica a feature de maior impacto. Ao contrário do observado para a Formação Barão de Guaicuí, K é a variável de menor importância.

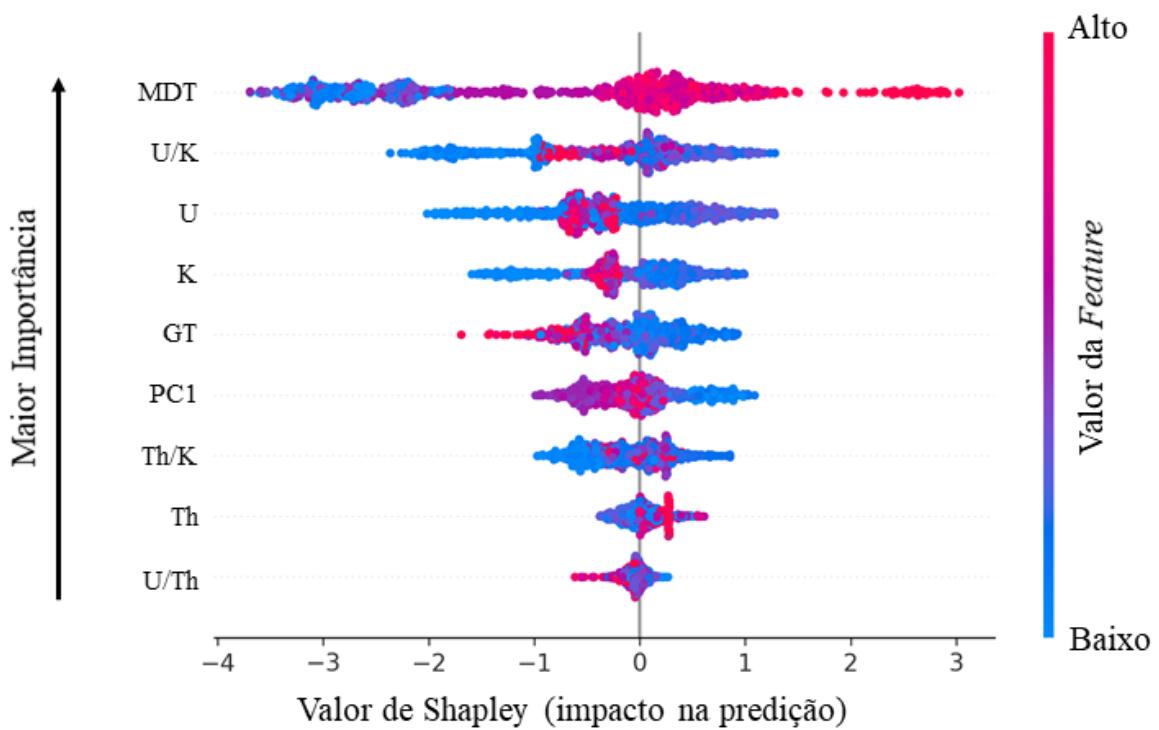


Figura 48 - Summary plot da Formação Bandeirinha.

A Figura 50 mostra que valores baixos de U e U/Th levam o modelo a predizer a Formação Sopa-Brumadinho (PP4esb), ao passo que valores elevados dessas variáveis desfavorecem a classificação dessa unidade.

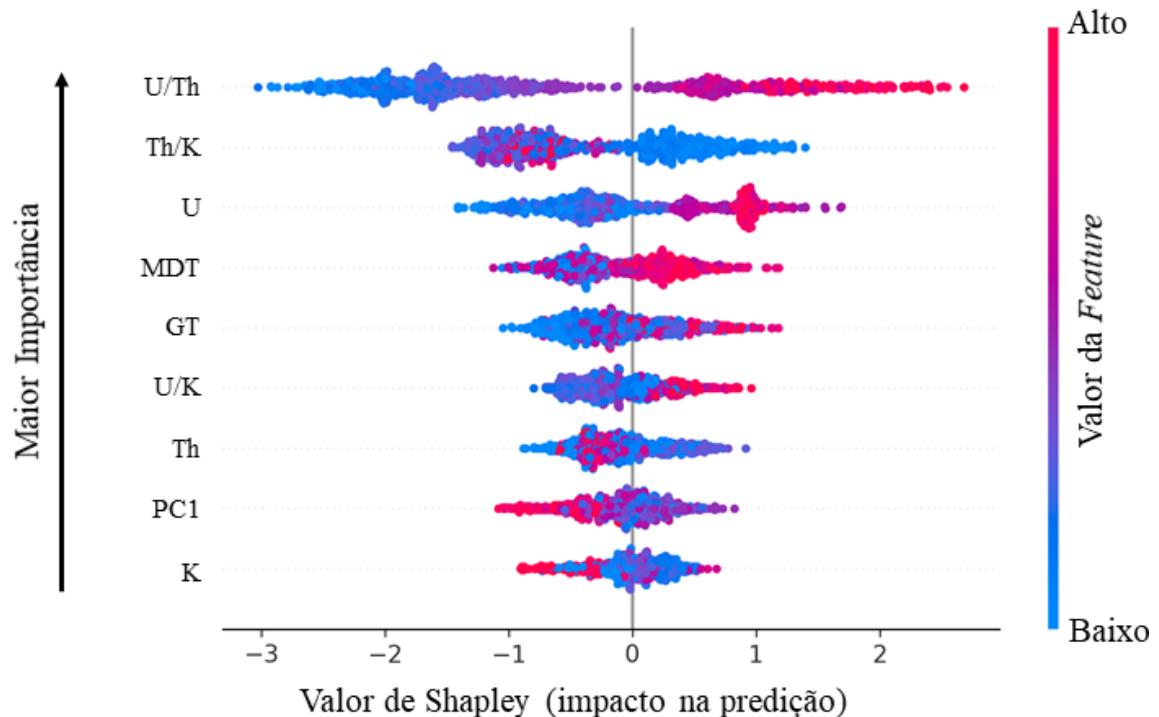


Figura 49 - Summary plot da Formação São João da Chapada.

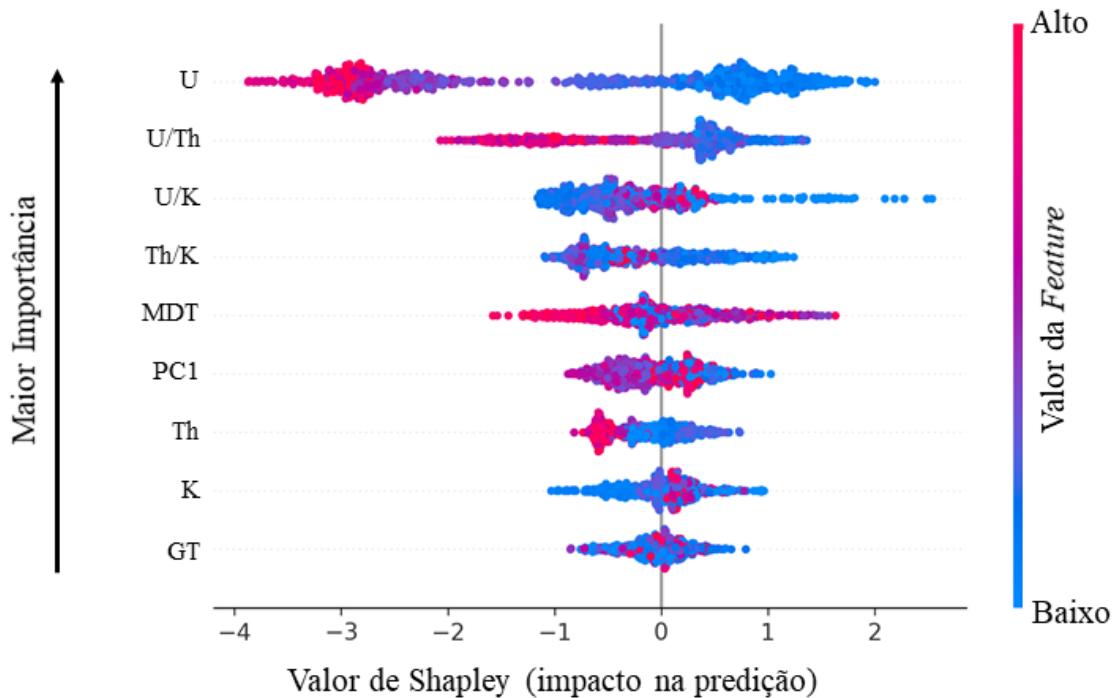


Figura 50 - Summary plot da Formação Sopa-Brumadinho.

A partir da observação da Figura 51, é possível perceber que baixos valores dos canais radiométricos (*i.e.* U, Th e K) impactam positivamente na predição de uma instância como Formação Galho do Miguel (PP4egm). Além disso, nota-se que valores muito baixos de elevação (MDT) também levam o modelo a predizer essa unidade.

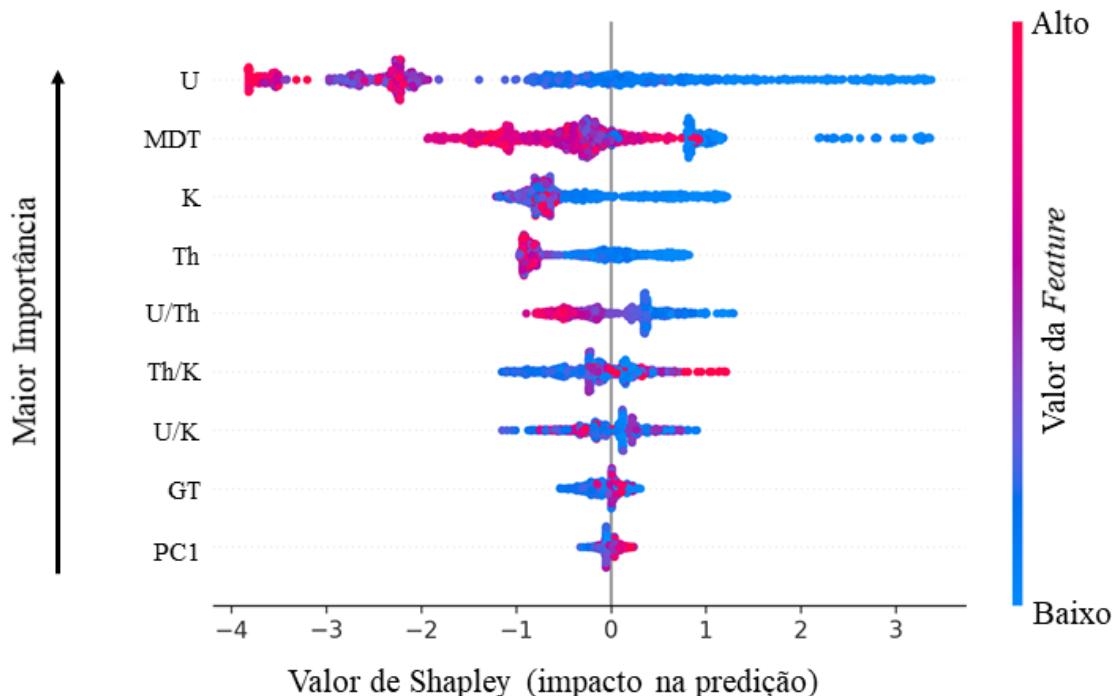


Figura 51 - Summary plot da Formação Galho do Miguel.

6.9 ANÁLISE DE FENÔMENOS ASSOCIADOS A DADOS GEOESPACIAIS

A Figura 52 apresenta as distribuições bivariadas dos canais radiométricos agrupadas nos conjuntos de treino e teste. Essas variáveis geofísicas encontram-se estandardizadas. De forma geral, nota-se que não há distorções significativas nas distribuições bivariadas quando subdivididas nos conjuntos de treino e teste.

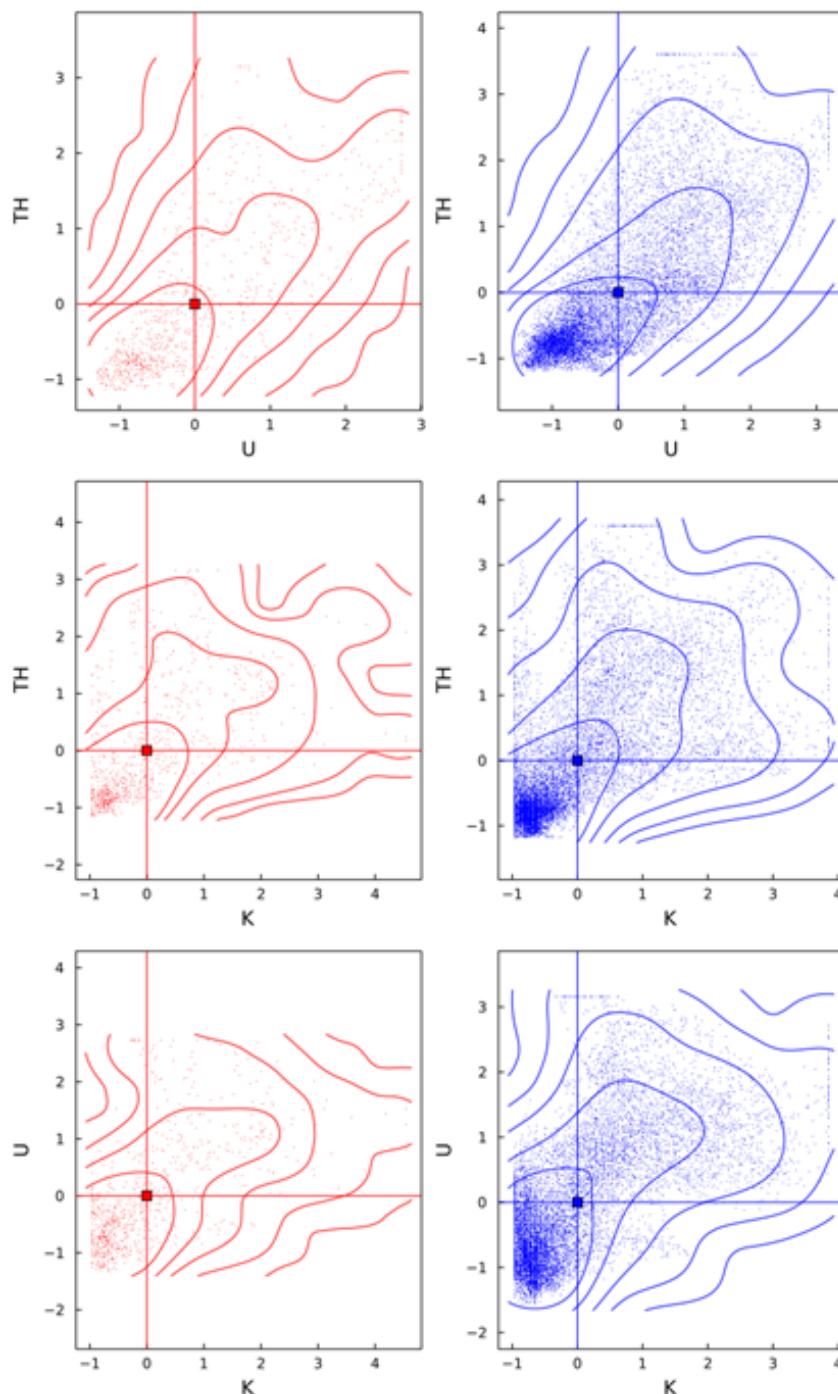


Figura 52 - Distribuições bivariadas dos canais radiométricos agrupadas pelos conjuntos de treino (vermelho) e teste (azul). Essas variáveis apresentam-se estandardizadas.

A Figura 53 mostra os variogramas experimentais N-S (000°-180°) das variáveis radiométricas K, Th e U. O eixo horizontal representa a distância entre pares de amostras, enquanto o eixo vertical representa a variância espacial. A partir da análise desses variogramas, nota-se a clara existência de estrutura espacial em todos os canais radiométricos, com alcances que variam entre 1.500 e 3.000 metros.

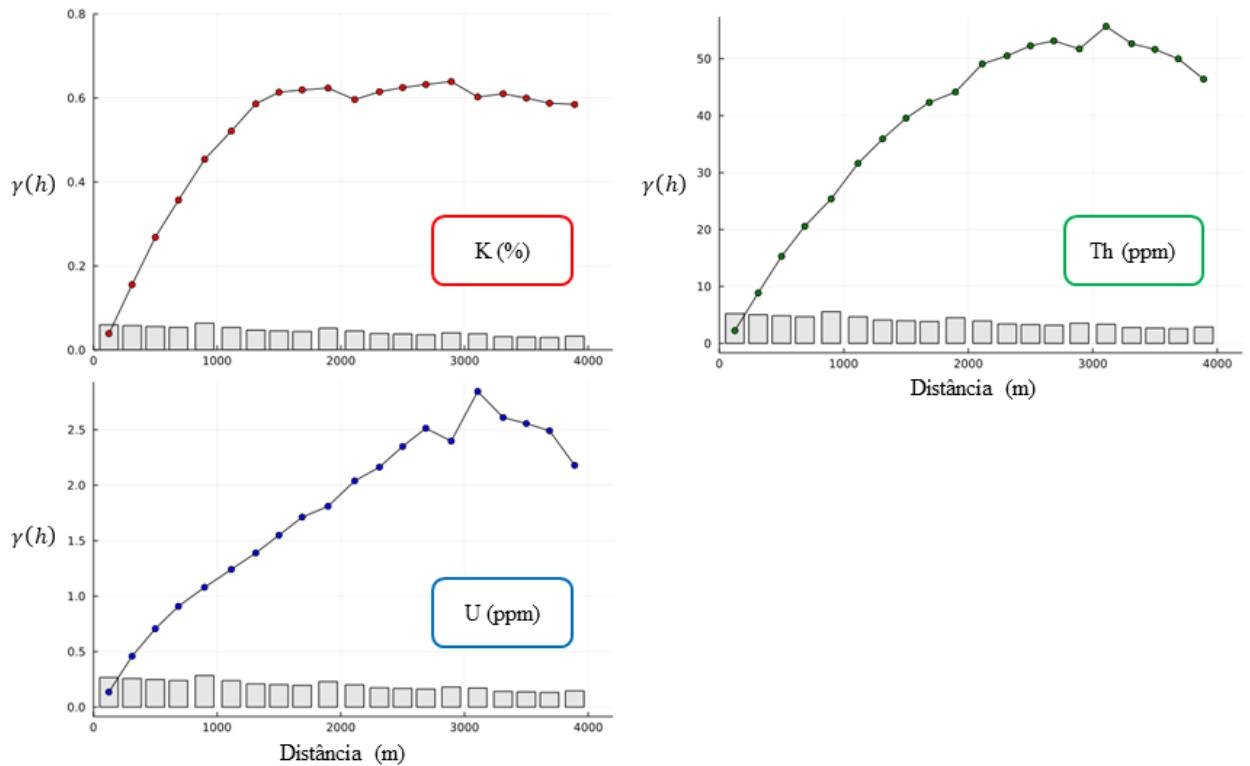


Figura 53 - Variogramas experimentais N-S das variáveis K (vermelho), Th (verde) e U (azul).

7 DISCUSSÃO

7.1 MAPAS GEOLÓGICOS PREDITIVOS

7.1.1 Seleção do Modelo de Melhor Performance

Dentre os oito algoritmos de Aprendizado de Máquina avaliados, os modelos de EL (*i.e.* *Random Forests* e *XGBoost*) apresentaram os maiores *scores* de VCKF para as métricas *F1-score*, *recall*, precisão e acurácia (Figura 38). CRACKNELL & READING (2014) e COSTA *et al.* (2019) também relataram que o algoritmo *Random Forests*, na confecção de mapas geológicos preditivos, mostrou performances superiores a modelos como *Naive Bayes*, *K-Nearest Neighbors*, *Support Vector Machine* e *Multilayer Perceptrons*, embora não tenham avaliado o modelo *XGBoost*.

Entretanto, não se pode descartar que a performance do modelo *Multilayer Perceptrons* tenha sido inferior a dos modelos de EL em função do elevado custo computacional para otimizar seus diversos hiperparâmetros, em especial, o número de camadas ocultas e o número de neurônios que as compõem⁵³. Uma alternativa seria a utilização de estratégias de Aprendizado de Máquina Automatizado, já que uma das suas principais tarefas é justamente a otimização automática dos hiperparâmetros de modelos complexos a um custo computacional relativamente inferior (HUTTER *et al.*, 2019).

Como os *scores* de VCKF para os modelos de EL são muito similares entre si (*i.e.* diferenças de 0,01), não foi possível definir qual desses dois algoritmos apresentou a melhor performance apenas com os resultados das métricas. Dessa forma, outras abordagens foram conduzidas para compará-los, sendo elas os mapas de probabilidade por unidade litoestratigráfica e os mapas de entropia da informação. A partir da comparação entre as distribuições espaciais de entropia (Figura 44), nota-se que o *XGBoost* gerou previsões com menor incerteza associada em relação ao modelo *Random Forests*. Os mapas de probabilidade por classe (Figura 42 e Figura 43) também corroboram com essa observação, já que as probabilidades vinculadas ao *XGBoost* evidenciam uma maior confiabilidade nas previsões. Em outras palavras, quando se compara as probabilidades preditas para uma determinada unidade, percebe-se que o *XGBoost* exibe maiores valores (*i.e.* cores mais quentes) do que o modelo *Random Forests*.

⁵³ Parâmetro *hidden_layer_sizes* do framework Scikit-Learn.

Portanto, com base nos *scores* de VCKF e nos mapas de probabilidade por classe e de entropia, o modelo *XGBoost* foi selecionado como o mais adequado para a confecção do mapa geológico preditivo final da área deste estudo (Anexo IV).

7.1.2 Análise das Predições das Unidades Litoestratigráficas

A partir da matriz de confusão do conjunto T_b (Figura 40) e dos mapas de erros de classificação (Figura 41), nota-se que a maior fonte de inconsistência nos mapas preditivos está associada à dificuldade de diferenciação entre as formações São João da Chapada e Sopa-Brumadinho. Isso pode ser explicado pela intensa subamostragem dessas unidades durante a separação entre T_a e T_b , tendo em vista que ambas abrangem a maior parte da área segundo o mapa geológico integrado (Figura 11). Outro possível motivo seria a semelhança litológica entre ambas as unidades, uma vez que são essencialmente constituídas por quartzitos e filitos e, subordinadamente, por níveis conglomeráticos (KNAUER, 2007).

Por outro lado, os modelos encontraram menos dificuldade na predição das unidades Complexo Granito-Gnáissico e Formação Galho do Miguel, o que pode ser evidenciado pela Figura 40. Durante a construção do conjunto T_a , essas classes não foram submetidas a uma subamostragem tão severa, uma vez que representam unidades espacialmente restritas na área. No caso do Complexo Granito-Gnáissico, o alto número de VP também pode ser explicado pela grande discrepância litológica dessa unidade quando comparada com as demais, uma vez que ela é majoritariamente constituída por granitoides e gnaisses (CHAVES *et al.* 2012). A menor dificuldade na predição da Formação Galho do Miguel, por sua vez, pode ser justificada pela homogeneidade litológica dessa unidade, já que 90% dela é composta por quartzitos puros (KNAUER, 2007).

A comparação entre os mapas geológicos integrado e preditivo (Figura 54) permite a identificação de áreas potenciais para o refinamento do mapeamento geológico. Na porção centro-oeste, o mapa geológico integrado mostra uma ocorrência exclusiva da Formação São João da Chapada. Entretanto, todos os mapas preditivos sugerem ocorrências de outras unidades, como as formações Bandeirinha, Sopa-Brumadinho e Galho do Miguel. Já na região central, o mapa integrado apresenta uma falha normal de direção NW-SE com um pequeno deslocamento em planta do teto para NW. Todavia, a geometria da Formação Barão de Guaicuí, nos mapas preditivos, sugere um deslocamento em planta de direção W-E. Além disso, na região sudeste, o mapeamento realizado indica a predominância da Formação São João da Chapada em contato com a Formação Barão de Guaicuí. Entretanto, todos os mapas

preditivos mostram que essa área também apresenta similaridades com a Formação Bandeirinha. Adicionalmente, os mapas de entropia (Figura 44) exibem elevados valores de incerteza para essa região, enfatizando a necessidade de um refinamento do mapeamento.

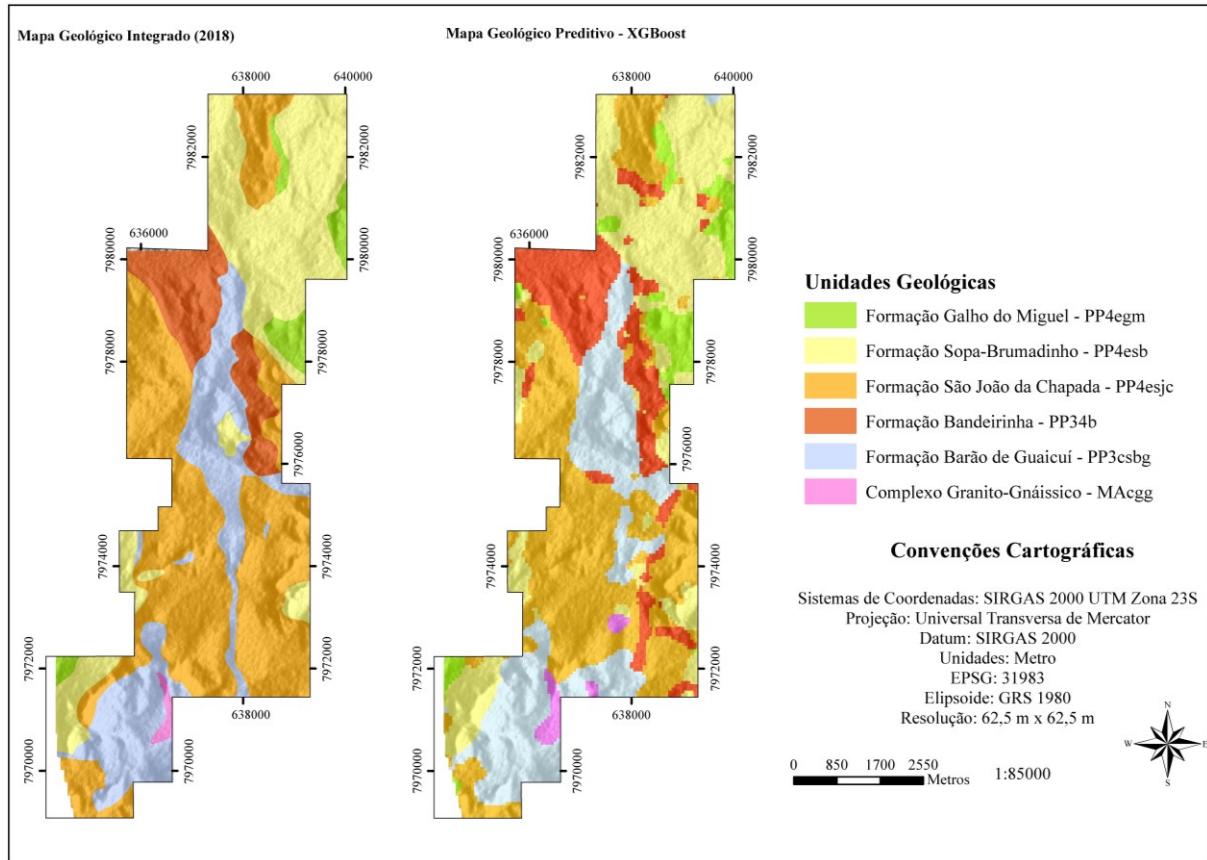


Figura 54 - Comparação entre os mapas geológicos integrado (esquerda) e preditivo (direita).

7.1.3 Desempenho da Validação Cruzada K-Fold na Seleção do Modelo

Recentemente, HOFFIMANN *et al.* (2021) discutiram limitações da utilização de técnicas clássicas de reamostragem (*e.g.* VCKF) na seleção do modelo de melhor performance em problemas geoespaciais. Nesse sentido, em função da ocorrência de fenômenos tipicamente relacionados a dados regionalizados (*e.g.* correlação espacial e distorção da distribuição bivariada), a VCKF tende a gerar estimativas do erro de generalização superotimistas e, consequentemente, elencar os modelos de forma inadequada.

Os variogramas dos canais radiométricas (Figura 53) evidenciam a presença de correlação espacial, enquanto os diagramas de dispersão das mesmas variáveis (Figura 52) sugerem uma ausência ou distorção mínima da distribuição bivariada entre os conjuntos de treino e teste. A inexistência (ou presença limitada) desse *shift* bivariado pode ser explicada

pelo fato de os conjuntos de treino e teste terem sido amostrados de uma mesma população e se situarem na mesma área geográfica. No trabalho de HOFFIMANN *et al.* (2021), a distorção da distribuição bivariada foi observada apenas quando as amostras de treino e teste se localizavam em áreas geograficamente distintas.

A partir da comparação entre os *scores* de VCKF e de teste (Figura 38), nota-se que a VCKF apresentou estimativas satisfatórias para a performance dos modelos (*i.e.* diferenças sempre inferiores a 0,10), ainda que ligeiramente otimistas em geral. Consequentemente, essa técnica de reamostragem obteve sucesso ao elencar os modelos em ordem decrescente de acordo com *F1-score*, conforme evidenciado pela Figura 39. Portanto, na presença de correlação espacial e na ausência de *shift* bivariado entre os conjuntos de treino e teste, a VCKF parece ser uma técnica adequada para a seleção do modelo de melhor performance. Essa conclusão também foi alcançada por HOFFIMANN *et al.* (2021), ainda que esses autores recomendem a utilização da técnica *Density Ratio Validation* como estimador do erro de generalização em problemas geoespaciais.

7.2 INTERPRETAÇÃO DO MODELO XGBOOST

Conforme discutido anteriormente, o modelo *XGBoost* foi selecionado como mais adequado de acordo com as métricas avaliadas e os mapas de probabilidade por unidade litoestratigráfica e de entropia da informação. Nesse sentido, uma análise foi conduzida com o intuito de interpretar as previsões geradas por esse modelo, utilizando o *framework* SHAP.

As variáveis que apresentaram maior impacto nas previsões do *XGBoost* foram os canais e razões radiométricos, enquanto a primeira componente principal, que representa a reprojeção das bandas Landsat 8, mostrou uma menor contribuição para as previsões (Figura 45). Uma constatação similar também foi relatada por CRACKNELL *et al.* (2014), KUHN *et al.* (2018) e COSTA *et al.* (2019) que, diferentemente, utilizaram outra técnica de avaliação do impacto das variáveis (*i.e.* Índice de Gini) e excluíram as bandas Landsat da etapa treinamento do algoritmo. A permanência das informações Landsat neste trabalho foi possível pela redução da alta correlação linear entre as bandas via ACP (Figura 36) e justificada pela limitada cobertura vegetal na área (GREBBY *et al.*, 2011).

O fato de altos valores dos canais radiométricos impactarem positivamente na previsão do Complexo Granito-Gnáissico (Figura 46) pode ser justificado pela composição félítica das rochas que compõem essa unidade. Segundo DENTITH & MUDGE (2014), o conteúdo de radioelementos em rochas ígneas tende a ser maior com o aumento do teor de sílica em

virtude da maior abundância de feldspatos e micas. Esses silicatos são representados, no Complexo Granito-Gnáissico, por K-feldspato, plagioclásio e biotita.

O grande impacto positivo da variável K na predição da Formação Barão de Guaicuí (Figura 47) se relaciona à abundância de minerais micáceos nessa unidade, como sericita e biotita. Nesse caso, a elevada concentração de K nessas rochas está associada à composição do protólito, uma vez que o metamorfismo não impacta significativamente o conteúdo de radioelementos (DENTITH & MUDGE, 2014). Ainda que a elevação não seja uma das *features* mais importantes, o fato de baixos valores de MDT favorecerem a predição da Formação Barão de Guaicuí pode estar vinculado à ocorrência dessa unidade em regiões escavadas por drenagens.

Altos valores de MDT influenciaram positivamente a predição da Formação Bandeirinha (Figura 48), indicando que essa unidade está posicionada, em média, em porções mais elevadas da paisagem regional, o que também é evidenciado pelos *boxplots* da Figura 24. Ainda que essa unidade seja constituída por quartzitos micáceos, a variável K não mostra um impacto tão evidente em sua predição. Isso possivelmente está relacionado ao fato de a quantidade de micas não ser suficiente a ponto de resultar respostas significativas de K, o que é evidenciado pelo sumário estatístico do Anexo III.

Diferentemente da abordagem adotada por COSTA *et al.* (2019), que excluiu as razões radiométricas do treinamento do algoritmo, neste trabalho, essas variáveis foram mantidas e apresentaram impactos significativos na predição de certas unidades. No caso da Formação São João da Chapada, por exemplo, altos valores de U/Th mostraram impactos positivos na predição dessa classe (Figura 49). Adicionalmente, o impacto positivo de altos valores de U e a baixa importância de Th sugerem um maior conteúdo em U em relação ao Th nos sedimentos que compõem as rochas da Formação São João da Chapada.

Assim como na Formação São João da Chapada, o U e a razão U/Th demonstraram uma forte influência na predição da Formação Sopa-Brumadinho (Figura 50). Entretanto, altos valores dessas variáveis impactam negativamente na predição dessa classe, ou seja, uma situação oposta àquela observada na unidade sotoposta. Portanto, os sedimentos formadores das rochas da Formação Sopa-Brumadinho possivelmente possuem um menor conteúdo em U do que em Th relativamente.

As predições da Formação Galho do Miguel foram positivamente impactadas por baixos valores dos canais radiométricos (Figura 51). Esse fato pode ser justificado já que 90% das

rochas que constituem essa unidade serem representadas por quartzitos puros (KNAUER, 2007). A resposta radiométrica de rochas sedimentares (e metassedimentares) é determinada predominantemente pela presença de feldspatos, micas e argilas (DENTITH & MUDGE, 2014) que, por sua vez, são minerais escassos nas rochas dessa classe. Além disso, o fato de baixos valores de MDT levarem o *XGBoost* a predizer a Formação Galho do Miguel aparenta ser contraintuitivo, já que ela representa a unidade de topo e é composta por litotipos relativamente resistentes a processos intempéricos. O padrão bimodal do histograma de MDT (Figura 31B) sugere que há uma mistura de populações e reforça a necessidade de uma reavaliação dessa unidade litoestratigráfica.

8 CONCLUSÃO

O mapeamento geológico representa uma etapa-chave para todos os empreendimentos que se relacionam, de alguma forma, à Geologia, sejam eles ligados à Exploração Mineral, ou a atividades ambientais, agrícolas e civis. É importante dinamizar a realização dessa atividade com o objetivo de ampliar a compreensão quantitativa das relações entre as unidades mapeadas. Sendo assim, a aplicação de técnicas de Aprendizado de Máquina Supervisionado mostra-se como uma ferramenta útil que busca, a partir de dados de sensores remotos, refinar e auditar mapas geológicos previamente confeccionados.

Dentre os oito algoritmos de Aprendizado de Máquina avaliados, os modelos *Random Forests* e *XGBoost* apresentaram as melhores performances com base na métrica *F1-score*, sendo a diferença entre eles ínfima. Os mapas de probabilidade por classe e de entropia da informação evidenciaram que o *XGBoost* gerou previsões com maior confiabilidade e menor incerteza associada do que o modelo *Random Forests*. Esses mapas também se mostraram importantes ferramentas para compreensão das incertezas associadas tanto ao mapa geológico pré-existente quanto ao mapa preditivo, além de salientarem regiões a serem reavaliadas. O modelo selecionado, *XGBoost*, aplicado pela primeira vez na confecção de mapas geológicos preditivos, apresentou resultados bastante satisfatórios para esse tipo de estudo.

A utilização do SHAP possibilitou um melhor entendimento das previsões realizadas pelo modelo selecionado e a relação delas com os sensores remotos. Pode-se dizer que o SHAP é uma importante ferramenta de validação geológica conceitual das previsões geradas por um modelo, permitindo a análise individual de cada unidade litoestratigráfica predita.

O desempenho da VCKF como técnica de seleção do modelo de melhor performance foi adequado na presença de correlação espacial (*i.e.* amostras não i.i.d.) e na ausência de distorção significativa das distribuições bivariadas entre os conjuntos de treino e teste. Portanto, em problemas de refinamento de mapas geológicos, a VCKF aparenta ser uma técnica adequada, desde que a heurística de separação entre treino e teste não gere *shifts* bivariados entre ambos os conjuntos.

A criação de um repositório remoto aberto que documenta e disponibiliza todo o fluxo de trabalho adotado estimula a aplicação de técnicas de Aprendizado de Máquina no mapeamento geológico. Novos estudos utilizaram este trabalho como referência, o que salienta ainda mais a necessidade e a importância do compartilhamento de conteúdo aberto à comunidade científica.

REFERÊNCIAS BIBLIOGRÁFICAS

- AHA, D. W. **Lazy learning**. New York: Springer Science & Business Media, 1997. 423 p.
- ALMEIDA ABREU, P. A. O Supergrupo Espinhaço da Serra do Espinhaço Meridional (Minas Gerais): O rifte, a bacia e o orógeno. **Geonomos**, v. 3, 1995. 1-18.
- ALMEIDA, F. F. M. O Cráton do São Francisco. **Revista Brasileira de Geociências**, v. 7, 1977. 349-364.
- ALKMIM, F. F.; PEDROSA-SOARES, A. C.; NOCE, C. M.; CRUZ, S. C. P. Sobre a evolução tectônica do Orógeno Araçuaí-Congo Ocidental. **Geonomos**, v. 15, 2007. 25-43.
- ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics Surveys**, v. 4, 2010. 40-79.
- BARUA, S.; ISLAM, M. M.; MURASE, K. A novel synthetic minority oversampling technique for imbalanced data set learning. In: **International Conference on Neural Information Processing**, 2011. 735-744.
- BERRY, M. W.; MOHAMED, A.; YAP, B. W. **Supervised and unsupervised learning for data science**. Switzerland: Springer Nature, 2020. 187 p.
- BEZANSON, J.; KARPINSKI, S.; SHAH, V. B.; EDELMAN, A. Julia: a fast dynamic language for technical computing. **arXiv preprint arXiv:1209.5145**, 2012. 1-27.
- BISHOP, C. M. **Pattern recognition**. New York: Springer Science & Business Media, 2006. 738 p.
- BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, 2001. 5-32.
- BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. California: Wadsworth & Brooks/Probability Series, 1984. 358 p.
- BURKOV, A. **The hundred-page machine learning book**. Canada: Andriy Burkov, 2019. 160 p.
- CEVIK, I. S.; LEUANGTHONG, O.; CATÉ, A.; ORTIZ, J. M. On the use of machine learning for mineral resource classification. **Mining, Metallurgy & Exploration**, v. 38, 2021. 2055-2073.

CHACON, S.; STRAUB, B. **Pro Git**. 2^a. ed. New York: Springer Science & Business Media, 2014. 426 p.

CHAVES, A. O; DUSSIN, T. M.; RENGER, F. E.; CHAVES, M. L. D. S. C. Petrografia e litogegeoquímica do magmatismo traqui-andesítico de Gouveia (MG): implicações genéticas e tectônicas. **Geociências (São Paulo)**, v. 31, n. 4, 2012. 504-515.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, 2002. 321-357.

CHEN, S. Interpretation of multi-label classification models using shapley values. **arXiv preprint arXiv:2104.10505**, 2021. 1-12.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 2016. 785-794.

CODEMIG – COMPANHIA DE DESENVOLVIMENTO ECONÔMICO DE MINAS GERAIS. **Levantamento Aerogeofísico de Minas Gerais, Área 04, Faixa Datas – São João da Chapada**. Magnetometria – Mapa de Sinal Analítico. Belo Horizonte, 2000/2001, em CD-ROM.

COSTA, I. S. L.; TAVARES, F. M.; DE OLIVEIRA, J. K. M. Predictive lithological mapping through machine learning methods: a case study in the Cinzento Lineament, Carajás Province, Brazil. **Journal of the Geological Survey of Brazil**, v. 2, n. 1, 2019. 26-36.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, 1967. 21-27.

CRACKNELL, M. **J. Machine learning for geological mapping: Algorithms and applications**. Tese de Doutorado. Tasmania: University of Tasmania, 2014. 275 p.

CRACKNELL, M. J.; READING, A. M. 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. **Computers & Geosciences**, v. 63, 2014. 22-33.

CRACKNELL, M. J.; READING, A. M.; MCNEILL, A. W. Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer–Mt Charter region, Tasmania, using Random

Forests™ and Self-Organising Maps. **Australian Journal of Earth Sciences**, v. 61, n. 2, 2014. 287-304.

CUI, K.; JING, X. Research on prediction model of geotechnical parameters based on BP neural network. **Neural Computing and Applications**, v. 31, n. 12, 2019. 8205-8215.

DE SOUZA FILHO, C. R.; CRÓSTA, A. P. Geotecnologias aplicadas à geologia. **Revista Brasileira de Geociências**, v. 33, n. 2, 2003. 1-4.

DEISENROTH, M. P.; FAISAL, A. A.; ONG, C. S. **Mathematics for machine learning**. New York: Cambridge University Press, 2020. 371 p.

DENTITH, M.; MUDGE, S. T. **Geophysics for the mineral exploration geoscientist**. New York: Cambridge University Press, 2014. 438 p.

DOMINGOS, P.; PAZZANI, M. On the optimality of the simple Bayesian classifier under zero-one loss. **Machine Learning**, v. 29, n. 2, 1997. 103-130.

DUTTA, S.; BANDOPADHYAY, S.; GANGULI, R.; MISRA, D. Machine learning algorithms and their application to ore reserve estimation of sparse and imprecise data. **Journal of Intelligent Learning Systems and Applications**, v. 2, n. 2, 2010. 86-96.

EL NAQA, I.; MURPHY, M. J. What is machine learning? In: **Machine Learning in Radiation Oncology**, 2015. 3-11.

FANTINEL, L. M. **O ensino de mapeamento geológico no Centro de Geologia Eschwege, Diamantina-MG. Análise de três décadas de práticas de campo (1970-2000)**. Tese de Doutorado. Campinas: Universidade Estadual de Campinas, 2005. 257 p.

FERRACIOLLI, M. A.; BOCCA, F. F.; RODRIGUES, L. H. A. Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models. **Computers and Electronics in Agriculture**, v. 161, 2019. 233-240.

FIGUEIRA, C. V. **Modelos de regressão logística. Dissertação de Mestrado**. Porto Alegre: Universidade Federal do Rio Grande do Sul, 2006. 138 p.

GREBBY, S.; NADEN, J.; CUNNINGHAM, D.; TANSEY, K. Integrating airborne multispectral imagery and airborne LiDAR data for enhanced lithological mapping in vegetated terrain. **Remote Sensing of Environment**, v. 115, n. 1, 2011. 214-226.

- GRIFFTH, D. A. **Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization**. Germany: Springer-Verlag, 2003. 260 p.
- GUYON, I. A practical guide to model selection. In: **Proceedings of the Machine Learning Summer School**, v. 11, 2009. 1-37.
- HARRIS, J. R.; GRUNSKY, E. C. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. **Computers & Geosciences**, v. 80, 2015. 9-25.
- HARVEY, A. S.; FOTOPOULOS, G. Geological mapping using machine learning algorithms. **International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences**, v. 41, 2016. 423-430.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of Statistical Learning: Data Mining, Inference and Prediction**. 2^a. ed. New York: Springer Series in Statistics, 2009. 533 p.
- HAYKIN, S. **Neural networks: a comprehensive foundation**. 2^a. ed. Delhi: Pearson Education, 1999. 823 p.
- HE, H.; BAI, Y.; GARCIA, E. A.; LI, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: **2008 IEEE international joint conference on neural networks**, 2008. 1322-1328.
- HEATON, J. An empirical analysis of feature engineering for predictive modeling. In: **SoutheastCon 2016 (IEEE)**, 2016. 1-6.
- HOFFIMANN, J.; ZORTEA, M.; DE CARVALHO, B.; ZADROZNY, B. Geostatistical Learning: challenges and opportunities. **arXiv preprint arXiv:2102.08791**, 2021. 1-28.
- HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. 2^a. ed. New York: John Wiley and Sons, 2000. 375 p.
- HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. **Automated machine learning: methods, systems, challenges**. Switzerland: Springer Nature, 2019. 219 p.
- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. New York: Springer, 2013. 426 p.

JOHN, G.; Langley, P. Estimating continuous distributions in Bayesian classifiers. In: **Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence**, 1995. 338-345.

KARPATNE, A.; EBERT-UPHOFF, S. R.; BABALE, H. A.; KUMAR, V. Machine learning for the geosciences: Challenges and opportunities. **IEEE Transactions on Knowledge and Data Engineering**, v. 31, n. 8, 2018. 1544-1554.

KEAREY, P.; BROOKS, M.; HILL, I. **Geofísica de exploração**. São Paulo: Oficina de Textos, 2009. 438 p.

KEOGH, E.; MUEEN, A. Curse of dimensionality. **Encyclopedia of Machine Learning and Data Mining**, 2017. 314-315.

KIM, K.; SHIM, P.; SHIN, S. An alternative bilinear interpolation method between spherical grids. **Atmosphere**, v. 10, n. 3, 123, 2019. 1-11.

CLUYVER, T.; RAGAN-KELLEY, B.; PÉREZ, F.; GRANGER, B. E.; BUSSONIER, M.; FREDERIC, J.; KELLEY, K; HAMRICK, J; GROUT, J.; CORLAY, S.; IVANOV, P.; AVILA, D.; ABDALLA, S.; WILLING, C. Jupyter Notebooks - a publishing format for reproducible computational workflows. In: **Proceedings of the 20th International Conference on Electronic Publishing**, 2016. 87-90.

KNAUER, L. G. **Evolução geologica do pré-cambriano da porção centro-leste da serra do Espigão meridional e metalogenese associada. Dissertação de Mestrado**. Campinas: Universidade Estadual de Campinas, 1990. 298 p.

KNAUER, L. G. O Supergrupo Espinhaço em Minas Gerais: considerações sobre sua estratigrafia e seu arranjo estrutural. **Geonomos**, v. 15, n. 1, 2007. 81-90.

KOTSIANTIS, S. B. Supervised machine learning: A review of classification techniques. **Emerging artificial intelligence applications in computer engineering**, v. 160, n. 1, 2007. 3-24.

KUHN, M.; JOHNSON, K. **Applied predictive modeling**. New York: Springer, 2013. 600 p.

KUHN, M.; JOHNSON, K. **Feature engineering and selection: A practical approach for predictive models**. CRC Press, 2019. 297 p.

KUHN, S.; CRACKNELL, M. J.; READING, A. M. Lithological mapping via Random Forests: Information Entropy as a proxy for inaccuracy. **ASEG Extended Abstracts**, v. 2016, n. 1, 2016. 1-4.

KUHN, S.; CRACKNELL, M. J.; READING, A. M. Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. **Geophysics**, v. 83, n. 4, 2018. 183-193.

KUHN, S.; CRACKNELL, M. J.; READING, A. M. Lithological mapping in the Central African Copper Belt using Random Forests and clustering: Strategies for optimised results. **Ore Geology Reviews**, 112, 2019. 1-16.

KURT MENKE, G. I. S. P.; SMITH, R.; PIRELLI, L.; VAN HOESEN, G. I. S. P. **Mastering QGIS**. Packt Publishing Ltd., 2016. 388 p.

LAURIKKALA, J. Improving identification of difficult small classes by balancing class distribution. In: **Conference on Artificial Intelligence in Medicine in Europe**, 2001. 63-66.

LEVERINGTON, D. W. Discrimination of sedimentary lithologies using Hyperion and Landsat Thematic Mapper data: a case study at Melville Island, Canadian High Arctic. **International Journal of Remote Sensing**, v. 31, n. 1, 2010. 233-260.

LOPES-SILVA, L.; KNAUER, L. G. Posicionamento estratigráfico da Formação Bandeirinha na região de Diamantina, Minas Gerais: Grupo Costa Sena ou Supergrupo Espinhaço?. **Geonomos**, v. 19, 2011. 131-151.

LUNDBERG, S. M.; LEE, S. A unified approach to interpreting model predictions. In: **Proceedings of the 31st international conference on neural information processing systems**, 2017. 4768-4777.

MARTINS E SOUZA FILHO, P. W.; PARADELLA, W. R.; SOUZA JÚNIOR, C.; VALERIANO, D. D. M.; MIRANDA, F. P. D. Sensoriamento remoto e recursos naturais da Amazônia. **Ciência e Cultura**, 58(3), 2006. 37-41.

MATHERON, G. **The theory of regionalized variables and its applications**. École Nationale Supérieure des Mines de Paris, 1971. 211 p.

MEDEIROS, C. J. F.; COSTA, J. A. F. Uma comparação empírica de métodos de redução de dimensionalidade aplicados a visualização de dados. **Learning and Nonlinear Models - Revista da Sociedade Brasileira de Redes Neurais (SBRN)**, v. 6, n. 2, 2008. 81-110.

MICROSOFT INC. Documentation for Visual Studio Code. Disponível em: <<https://code.visualstudio.com/docs>>. Acesso em: 01 agosto 2021.

MINTY B. R. S. **A review of airborne gamma-ray spectrometric data-processing techniques**. Canberra: Australian Gov. Publ. Service, 1988. 48 p.

MITCHELL, T. **Machine learning**. McGraw-Hill: New York, 1997. 432 p.

NASA. Landsat 8 overview. Disponível em: <<https://landsat.gsfc.nasa.gov/landsat-8/landsat-8-overview>>. Acesso em: 15 julho 2021.

NOVO, E. M. M. **Sensoriamento remoto: princípios e aplicações**. 4^a. ed. São Paulo: Editora Blucher, 2010. 387 p.

PARAMOS FILHO, A. C.; MIOTO, C. L.; PESSI, D. D. ; GAMARRA, R. M.; DA SILVA, N. M.; CHAVES, J. R. **Geotecnologias para aplicações ambientais**. Maringá: Uniedusul Editora, 2021. 394 p.

PEDROSA-SOARES, A. C.; NOCE, C. M.; ALKMIM, F. F.; DA SILVA, L. C.; BABINSKI, M.; CORDANI, U.; CASTAÑEDA, C. Orógeno Araçuaí: síntese do conhecimento 30 anos após Almeida. **Geonomos**, v. 15, 2007. 1-16.

PFLUG, R. Observações sobre a estratigrafia da Série Minas na região de Diamantina, Minas Gerais. **DNPM/DGM, Not. Prel. Est. 142**, 1968. 1-20.

PURI, N.; PRASAD, H. D.; JAIN, A. Prediction of geotechnical parameters using machine learning techniques. **Procedia Computer Science**, v. 125, n. 1, 2018. 509-517.

RANDLES, B. M.; PASQUETTO, I. V; GOLSHAN, M. S; BORGMAN, C. L. Using the Jupyter notebook as a tool for open science: An empirical study. In: **2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)**, 2017. 1-2.

RASCHKA, S.; PATTERSON, J.; NOLET, C. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. **Information**, 193, v. 11, n. 4, 2020. 1-44.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why should i trust you?" Explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**, 2016. 1135-1144.

- RISH, I. An empirical study of the naive Bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**, 2001. 41-46.
- SAMSON, M. **Mineral resource estimates with machine learning and geostatistics. Dissertação de Mestrado**. Edmonton: University of Alberta, 2020. 99 p.
- SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of research and development**, v. 3, n. 3, 1959. 210-229.
- SANTOS, M. S.; SOARES, J. P.; ABREU, P. H.; ARAÚJO, H.; SANTOS, J. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches. **IEEE Computational Intelligence Magazine**, v. 13, n. 4, 2018. 59-76.
- SCHÖLL, W. U. Estratigrafia, sedimentologia e paleogeografia na região de Diamantina (Serra do Espinhaço, Minas Gerais, Brasil). **Forsch. Geol. Paläont.**, 51, 1980. 223-240.
- SCHÖLL, W.U.; FOGAÇA, A.C.C. Estratigrafia da Serra do Espinhaço na região de Diamantina. **Bol. Soc. Bras. Geol.**, 1, 1979. 55-71.
- SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. In: **International Conference on Machine Learning**, 2017. 3145-3153.
- SMOLA, A. J.; BARTLETT, P. J.; SCHUURMANS, D.; SCHÖLKOPF, B. **Advances in large margin classifiers**. MIT Press, 2000. 412 p.
- TALEBI, H.; PEETERS, L. J. M.; OTTO, A.; TOLOSANA-DELGADO, R. A truly spatial random forests algorithm for geoscience data analysis and modelling. **Mathematical Geosciences**, 2021. 1-22.
- TANG, A; CLARK, C. **ArcGIS® 9 geocoding rule base developer**. Redlands: ESRI, 2003. 167 p.
- TELFORD, W. M.; GELDART, L. P.; SHERIFF, R. E. **Applied geophysics**. 2^a. ed. New York: University of Cambridge, 1990. 770 p.
- THAKUR, A. **Approaching (almost) any machine learning problem**. Abhishek Thakur, 2020. 297 p.

THUNG, F.; BISSYANDE, T. F.; LO, D.; JIANG, L. Network structure of social coding in github. In: **2013 17th European conference on software maintenance and reengineering (IEEE)**, 2013. 323-326.

TUKEY, J. W. **Exploratory data analysis**. Princeton: Addison-Wesley Publishing Company, 1977. 506 p.

TURNER, J.; CHARNIAK, E. Supervised and unsupervised learning for sentence compression. In: **Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)**, 2005. 290-297.

USGS. Landsat 8. Disponível em: <<https://www.usgs.gov/core-science-systems/nli/landsat/landsat-8>>. Acesso em: 15 julho 2021.

VAN DER MAATEN, L.; POSTMA, E.; VAN DEN HERIK, J. Dimensionality reduction: a comparative. **Journal of Machine Learning Research**, v. 10, n. 13, 2009. 1-35.

VAN ROSSUM, G. Python Programming Language. In: **USENIX Annual Technical Conference**, v. 41, n. 1, 2007. 1-36.

VAPNIK, V. **The nature of statistical learning theory**. New York: Springer Science & Business Media, 1995. 188 p.

VISA, S.; RAMSEY, B.; RALESCU, A. L.; VAN DER KNAAP, E. Confusion matrix-based feature selection. **MAICS**, v. 710, n. 1, 2011. 120-127.

WEISS, G. M.; PROVOST, F. The effect of class distribution on classifier learning. **Rutgers University Library (Online Resource)**, 2001. 1-6.

WELLMANN, J. F.; REGENAUER-LIEB, K. Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models. **Tectonophysics**, v. 526, 2012. 207-216.

YANG, L.; SHAMI, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. **Neurocomputing**, v. 415, 2020. 295-316.

ZHOU, X.; ABEL, D. J.; TRUFFET, D. Data partitioning for parallel spatial join processing. **Geoinformatica**, v. 2, n. 2, 1998. 175-204.

ZÖLLER, M.; HUBER, M. F. Benchmark and survey of automated machine learning frameworks. **Journal of Artificial Intelligence Research**, v. 70, 2021. 409-472.