Instituto de Geociências

Universidade de São Paulo

Trabalho de Formatura - 21/46

Análise de Risco-Retorno em mapeamento de potencial mineral usando *machine learning* na Província Mineral de Alta Floresta

Aluno: Victor Silva dos Santos

Orientador: Vinicius Hector Abud Louro

Coorientador: Erwan Gloaguen

São Paulo

Dezembro 2021

Dedico este trabalho às pessoas das quais
foi tomado o direito de estudar.

Agradecimentos

Agradeço a todos, todas e todes que passaram pela minha vida e acreditaram que eu poderia escolher quem seria. Agradeço aos meus pais, em especial à minha mãe, Rosalia, uma mulher com o coração do tamanho do mundo que me criou com todo o cuidado e carinho e a quem devo tudo, À minha avó, Elzi, mulher forte que me colocou limites e me ensinou a dar valor às pequenas coisas da vida. Às minhas tias Mi, Tal, Il, Lena, Sé, Lia, Ana, Jô, que foram muito importantes na minha criação e me mostram a potência da feminilidade. Agradeço aos meus tios, primos, primas e afilhado desta gigante família mineiro-baiana: Carol, Gabriel, Marquinhos, Andi, Gabi, Jhey, Jheimy, Rilton, Babi, Jéssica, Du, Gi, Dessa, João Vitor, Arthur, Jefferson, Valdson, Alê, Ju, Joma, Edmundo, João e demais parentes.

Pelos quase 5 anos de amor, companheirismo e crescimento, agradeço à minha companheira Mariana, quem muito me apoiou durante o curso e fez dos últimos anos um dos melhores período da vida. À Dani e ao resto da família da minha companheira, agradeço pelas ótimas conversas e cuidado. Aos meus amigos, amigas e amiges, Waze, Sizu, Nit, Brisão, Lilo, Ocotô, Trevas, Kibe, Menoseu, Bassan, Rico, Tom, Dino, Grosa, Jura, Bariri, Tey, Mônica, Naru, Erílio, Etienne, Amin, Rê, Cath, Isa, Willy, Maris e demais pessoas que a vida me trouxe, obrigado pelos inesquecíveis momentos, pelo carinho, aprendizado e por serem pessoas incríveis na minha vida. Agradeço à galera do Cursinho da FFLCH, GGEO e Geojúnior pelo crescimento proporcionado. Pelo papel que tiveram na minha evolução profissional, agradeço à AnaMaX Contabilidade, Marcenaria Planeje e ao pessoal do IPT, INRS e Geoscan.

Agradeço às professoras, professores e educadores que dedicaram muita energia, tempo e cuidado e foram fundamentais na minha formação como indivíduo. Em especial, agradeço à Marilene, pelas ajudas, conselhos e por apontar caminhos, aos professores e professoras do Cursinho da FEA e MED, ao Boggi, pelas piadas e por deixar o curso mais leve, ao LAPS e Erwan, por despertarem minha paixão pelas geociências, e ao meu orientador Vinicius Louro, quem muito me apoiou durante a minha vida acadêmica e topou se lançar em novas áreas do conhecimento comigo.

Por tornarem os trabalhos, estudos e viagens de campo possíveis, agradeço aos funcionários e funcionárias do IGc-USP, Esthelinha, Samuca, Bira e demais.


HEY, GRAMPA, LOOK AT THIS COOL ROCK!
YES, THAT'S AN IGNEOUS ROCK. THAT MEANS IT WAS FORMED FROM MOLTEN LAVA.
I'M KIND OF AN EXPERT ON ROCKS AND MINERALS.
YOU SEE THAT BIG BOULDER UPON THE HILL? THAT'S IGNEOUS ROCK TOO. PROBABLY GRANITE.
WOW, ARE YOU SURE? HOW COME IT'S MOVING?
OR IT COULD BE A COW. I'M KIND OF AN EXPERT ON THEM, TOO.
8/31
BRIAN CRANE

Resumo

A exploração mineral é um campo crucial e estratégico para o desenvolvimento econômico-tecnológico e o bem-estar social. Apesar da crescente demanda por recursos minerais, descobertas globalmente relevantes têm se tornado cada vez mais raras uma vez que os depósitos rasos e/ou mais óbvios se tornaram escassos. Ferramentas que utilizam o aprendizado de máquina (machine learning) têm apresentado grande potencial, auxiliando geocientistas na solução de problemas em todos os campos, incluindo a descoberta de depósitos minerais. A Província Mineral de Alta Floresta (Mato Grosso - MT), vive um novo ciclo exploratório e mostra potencial mineral ainda não plenamente compreendido. Conta com dados aerogeofísicos, geológicos e geoquímicos em densidade e qualidade satisfatórias para o emprego de técnicas de *machine learning*. Neste projeto foi conduzido o mapeamento de potencial mineral da Província Mineral de Alta Floresta (PMAF) usando métodos de *machine learning* (ML), a avaliação da explicabilidade e performance dos modelos treinados e a quantificação de incertezas inerentes a este processo. Dados gamaespectrométricos, magnetométricos, modelos digital de elevação e geoquímica de sedimento de corrente foram utilizados na geração de vetores de exploração que representem fonte, transporte e alterações relacionadas aos depósitos auríferos da província. Mapas de gradiente de magnetismo máximo, distância até lineamentos magnéticos, razões Th/K e U/K, elevação e associações Au-Cu foram derivados a partir dos dados brutos visando modelar as assinaturas dos depósitos da região. Mil modelos de ML foram computados para simular possíveis cenários de favorabilidade aurífera utilizando os algoritmos random forest, gradient boosting, support vector machine e k-nearest neighbors. A área de estudo foi divida em setores norte e sul, sendo as amostras o primeiro utilizadas para treinamento e do segundo para teste. Com base na média e variância das predições obtidas foi possível definir áreas onde os modelos foram mais ou menos estáveis e fazer interpretações pautadas no modelo metalogenético da PMAF. As regiões com presença de depósitos e/ou ocorrências minerais conhecidas foram, de modo geral, classificadas como favoráveis à presença de depósitos de ouro. Dentre os produtos deste trabalho, foi produzido um manuscrito que será submetido em revistas internacionais.

Conteúdo

1 Introdução **6**

2 Objetivos **7**

3 Justificativa **7**

4 Materiais e Métodos **7**

4.1 Base de dados 7

4.2 Ferramentas utilizadas 8

5 Resultados obtidos **8**

5.1 "Risk-Return analysis in mineral prospectivity mapping using machine learning methods: A case study from Alta Floresta Mineral Province, northern Mato Grosso, Brazil" 8

6 Considerações finais **28**

7 Bibliografia **29**

1 Introdução

O mapeamento de potencial mineral (MPM) consiste em um processo de multiplos critérios que inclui a análise e síntese de dados geocientíficos. Este dados devem representar vetores de exploração capazes orientar a tomada de decisão na definição de novas áreas prospectivas para um dado minério (Carranza, 2015). Trata-se de uma metodologia computacional capaz de delimitar as regiões com maiores probabilidades de conterem depósitos minerais de um tipo particular. O MPM pode ser guiado por conhecimento geológico e pela compreensão dos mecanismos formadores do minério (*knowledge-driven*), ou pela informação/assinatura presente em dados, usada como vetor de exploração do minério (*data-driven*). A metodologia pode ainda se desenvolver de modo híbrido, pela utilização das duas abordagens (Carranza, 2008; Bonham-Carter, 2014).

Embora a descoberta de depósitos minerais constitua uma tarefa complexa, uma vez que se tratam do resultado de uma variedade de processos geológicos interdependentes em diferentes escalas no tempo e espaço, seus processos formadores costumam produzir assinaturas e evidências que guiam os profissionais nas etapas de exploração. Tradicionalmente, a definição de áreas com potencial para mineração ocorre através da análise de dados coletados em campo, por aerolevantamentos e por satélites, cartografados ou integrados em ambientes de Sistema de Informação Geográfica (SIG). No entanto, tratam-se de metodologias limitadas sob o ponto de vista de análises multivariadas e do estabelecimento relações não-lineares entre vetores de exploração (Porwal et al., 2003).

Os métodos de aprendizado de máquina têm trazido melhoras significativas ao MPM e representado uma ferramenta poderosa na criação de modelos que classificam regiões favoráveis a mineralizações (e.g., Sun et al., 2019). Também permitem a modelagem da processos complexos e não-lineares, capazes de capturar relações entre variáveis ou vetores de exploração até então impossíveis com o uso de métodos tradicionais. Além disso, possibilitam a abordagem de incertezas inerentes o processo de MPM (e.g., efeito de amostras negativas no treinamento de modelos) bem como análises de Risco-Retorno (Zuo, 2020), que podem auxíliar no processo de tomada de decisão.

Situada na porção sul do Cráton Amazônico, norte do Estado do Mato Grosso, a Província Mineral de Alta Floresta (PMAF) é um dos focos da exploração por ouro no Brasil. A província destaca-se por sua relevante produção nas décadas de 1970-1990 (160 t) e, mais recentemente, por descobertas de depósitos do tipo orogenético, epitermal e pórfiro (Barros, 1994; Assis, 2011; Trevisan, 2015; Teixeira, 2015). Conta com uma das maiores densidades de dados públicos do país, incluindo levantamentos aerogeofísicos, campanhas de geoquímica de sedimentos de corrente e mapeamentos geológicos em escalas a partir de 1:500.000, a configurando como um estudo de caso adequado para a aplicação de métodos de *machine learning*.

Este projeto foi desenvolvido em continuação às atividades realizados pelo aluno durante o projeto PIBIC(2019/2157) e estágio no Institut National de la Recherche Scientifique (INRS), Québec, Canadá, financiado pela Agência de Inovação da Universidade de São Paulo (AUSPIN) desenvolvido ao longo do ano de 2020.

2 Objetivos

Neste projeto, o objetivo central foi avaliar do potencial de métodos de *machine learning* na caracterização do potencial aurífero de novas áreas da PMAF. Além deste, se buscou: i) melhorar a explicabilidade dos modelos de ML para análise da congruência entre os resultados dos modelos de ML e os modelos genéticos dos depósitos da PMAF; ii) quantificar incertezas associadas à seleção de amostras negativas durante o treinamento de modelos de ML em MPM; iii) gerar mapas de potencial aurífero e risco-retorno de parte do setor leste da PMAF; iv) produzir um manuscrito em inglês para submissão em revistas internacionais.

3 Justificativa

Os recursos naturais servem de alicerce para o desenvolvimento tecnológico de diversos setores sociais e, atualmente, são essenciais ao desafio de transição para energias verdes ou renováveis. Entretanto, descobertas de depósitos relevantes em regiões rasas e óbvias são cada vez mais escassas, exigindo o desenvolvimento e aplicação de novas estratégias capazes de extrair mais informação a partir de dados geocientíficos e de gerar resultados mais confiáveis. Visando suprir tais demandas, o aprendizado de máquina se destaca por ser capaz de extrair valor de dados já existentes e de sintetizar processos complexos formadores de minério com o auxílio de geocientistas (Caté et al., 2017). Nesse contexto, a PMAF, com seu potencial ainda não plenamente compreendido (Assis et al., 2012; Junior et al., 2013; Silva et al., 2014; Assis, 2015; Fernandes et al., 2018), qualidade e densidade de dados aerogeofísicos, geológicos e geoquímicos satisfatórias ao emprego de métodos de ML, se caracteriza como um caso de estudo adequado para o escopo deste projeto.

4 Materiais e Métodos

4.1 Base de dados

O projeto utilizou dados públicos, em sua maioria dispobilizados na base de dados GeoSGB, do Serviço Geológico do Brasil, contando com dados aerogeofísicos (magnéticos e gamaespectrométricos), geoquímicos (sedimento de corrente em quadrantes de 15 a 5 km^2), mapas geológicos e localidades de ocorrências minerais e depósitos auríferos primários. Também foram utilizados dados da Missão Topográfica Radar Shuttle (SRTM) e disponíveis na base de dados *Earth Explorer*, do Serviço Geológico dos Estados Unidos (USGS). Os métodos utilizados neste trabalho encontram-se melhor descritos ao longo do manuscrito em inglês (Seção 5).

4.2 Ferramentas utilizadas

O projeto foi desenvolvido através do uso do *softwares* livres de Sistemas de Informação Geográfica (SIG), QGIS (QGIS Development Team, 2009), nas etapas de definição de lineamentos magnéticos e interpretações geológicas cabíveis. Para análise, visualização e processamento de dados foi utilizada a linguagem de programação Python (Van Rossum, 1995), contando com as bibliotecas: Numpy, Matplotlib, Pandas, Geopandas, Scipy, Fatiando a Terra, Geostatspy. No desenvolvimento de *workflows* e aplicação de algoritmos de *machine learning*, foi utilizada a biblioteca Scikit-Learn.

5 Resultados obtidos

5.1 "Risk-Return analysis in mineral prospectivity mapping using machine learning methods: A case study from Alta Floresta Mineral Province, northern Mato Grosso, Brazil"

Abstract

Mineral prospectivity/potential mapping using machine learning methods involves multiple uncertainties. Using negative random sampling and establishing likely sites to find new deposits commonly separate prospective from non-prospective zones during the training and produce various effects on predicted locations of potential maps. These effects include variation in the probabilistic outputs as well as in the contribution of the features. In this paper, we quantified the uncertainty associated with such factors by computing multiple scenarios. A risk-return analysis is conducted on the predictions of four machine learning algorithms, including Random Forest, Gradient Boosting, Support Vector Machine, and K-Nearest Neighbors. The algorithms performed a thousand iterations, producing four end maps with the mean and variance of the probabilistic outputs and risk-return domains. The Alta Floresta Mineral Province, Mato Grosso State, Brazil, was chosen as a case study to illustrate this methodology. .

Keywords: Mineral prospectivity mapping, Machine learning, Risk-return analysis, Alta Floresta Mineral Province, GIS

1 Introduction

In mineral prospectivity mapping (MPM), the goal is to define favorable areas to find new deposits. It consists of a multi-criteria process that can be guided by geological knowledge, i.e., by comprehension of the responsible mechanisms for ore-formation (knowledge-driven), by geoscientific data that can serve as signature or represent a ore-forming setting, able to drive the search for new potential deposit areas (data-driven), or a hybrid approach (knowledge- and data-driven)(Carranza, 2015).

Traditionally, the delineation of mineral exploration targets is conducted through the analysis of field, airborne, or satellite data on maps or in Geographic Information System (GIS). Manual data processing is affected by the limited capacity of the human brain to process multiple variables at the same time. Consequently, leads to lack of reproductivity, depending only on the assumption made by the geologist. Recently, Machine Learning (ML) algorithms have been used in order to overcome this limitations by automatically mapping relationships in n-dimensional datasets. Within the ensemble of ML algorithms, supervised learning methods may drive to more accurate prospective zones delimitation, once they can be used to find the underlying models by learning relationships from evidential layers to classify probable mineralized locations (e.g., Zuo and Carranza, 2011; Rodriguez-Galiano et al., 2015). In MPM, supervised approaches consist of dichotomous problems in which, in order to classify whether a location is prospective or not, models must be provided with two classes: i) a class of interest (e.g., known deposit locations), ; ii) a contrasting class (e.g., non-deposit locations). Nevertheless, deposits are product of rare events and non-random ore-forming processes, whereas non-deposits are the result of common and random geological processes (Carranza, 2009), i.e., an uncertain class with unknown locations.

Many authors have studied spatial phenomena distribution and behavior through techniques such Fry analysis (Fry, 1979), fractal analysis (Mandelbrot, 1982) and point pattern analysis (Diggle, 1983; Gatrell et al., 1996). In regional- or district- scales, deposits constitute point entities whose distribution in space can be studied using such techniques, which corroborate the understanding of the association of mineral deposits with geological features that control their occurrence and geometry (Carranza, 2009). The spatial occurrences of a class of interest (positive samples) support selecting contrasting examples (negative samples) to its behavior. However, this process maintains a degree of uncertainty because (i) even studying spatial patterns, it is still difficult to map all scales at which mineralization controls operate. Thus, it is unknown which minimum distance from the deposits could guarantee the selection of negative samples in non-prospective locations. (ii) It is unfeasible to confirm that all randomly picked points, constrained by given criteria, represent negative samples. (iii) Data are commonly available at different scales, having to be interpolated to generate predictive maps, so there is also a distinct uncertainty related to the evidential feature values at each location.

Previous works approached negative training samples creation using spatial pattern criteria (Carranza et al., 2008). These researches propose some ways to obtain negative samples, including the selection of mineral deposits of a different type, random sampling constrained by geological knowledge, or the use of drilling sites that appear to be barren (Nykänen et al., 2015). Zuo, 2020, studied the effects of using random negative training samples, and Wang et al., 2020, applied a framework using Monte Carlo simulations to quantify the uncertainty associated to the integration of multi-resolution evidence layers in MPM.

In this paper, we performed a ML-based framework for risk-return analysis focusing on uncertainty propagation associated with random sampling and the boundary that separates prospective from non-prospective zones. We also analyzed their effects on the contribution of the features and on the performance of ML models. This methodology was applied to a real case study in Alta Floresta Mineral Province, northern Mato Grosso State, Brazil.

2 Geological setting

2.1 Regional geology and Tectonics

The Alta Floresta (or Juruena) Mineral Province (AFMP) is a polymetallic region located in the south-central portion of the Amazonian Craton (AC). The AC is a tectonically stable area of approximately 4.5 million km^2 composed of two Precambrian shields (Guiana and Central Brazil), which are separated in its central area by the Phanerozoic covers of the Solimões and Amazon basins (Almeida et al., 1981; Santos et al., 2000). Early studies divided the region into six major geochronological provinces: Central Amazonian ($>$ 2.3 Ga), Maroni-Itacaiúnas (1.95-1.80 Ga), Ventuari-Tapajós (1.95-1.80 Ga), Rio Negro-Juruena (1.8-1.55 Ga), Rondonian-San Ignácio (1.55-1.3 Ga), and Sunsás (1.3-1.0 Ga) (Tassinari and Macambira, 1999). These provinces evolved due to a sequence of continental accretion events characterized by a succession of magmatic arcs. Tassinari and Macambira, 1999, 2004, associate the origin of Sunsás, and some parts of Rondonian-San Ignácio and Maroni-Itacaiúnas provinces, with continental collisional events.

Santos, 2003; Santos et al., 2006, propose a scenario with seven tectonic provinces, including Carajás (3.0-2.5 Ga), Amazônia Central (Archean), Transamazonas (2.26-2.01 Ga), Tapajós-Parima (2.03-1.88 Ga), Rio Negro-Juruena (1.82-1.52 Ga), Rondônia-Juruena (1.82-1.4 Ga), and Sunsás-k'Mudku (1.45-1.10 Ga). Juliani et al., 2013 divide the south portion of the AC into two Paleoproterozoic magmatic arcs along the E-W direction.

According to Scandolara et al., 2017, the AFMP has been tectonically juxtaposed to the Tapajós Mineral Province (1.69 to 1.63 Ga) through a roughly E-W-trending shear zone. The combination of new results and observations gathered by Trevisan et al., 2021, together with pre-existing data from the literature on the plutonic and volcanic rocks of the eastern portion of the AFMP indicate long-lived magmatism in the southern part of the AC between 2037 and 1757 Ma. The authors introduced a modeled scenario where the region formed through tectonic switching represented by alternating periods of crustal thickening during flat subduction and crustal extension during slab-rollback at a single active convergent margin of an accretionary orogen.

2.2 The Alta Floresta Mineral Province and gold mineralization

The AFMP comprise a belt approximately 500 km long by 100 km wide constituted of Peixoto de Azevedo Domain (2.8-1.86 Ga), São Marcelo-Cabeça Group (1.85 Ga), Juruena Supersuite (1.81-1.75 Ga), Nova Monte Verde Complex (1.81-1.76 Ga), Colíder Group (1.81-1.76 Ga), Teles Pires suite (1.79-1.75 Ga), Roosevelt Group (1.76-1.74 Ga), Caiabis Group (1.98-1.37 Ga), and Alto Tapajós basin (0.30 Ga). The Peixoto de Azevedo Domain represents a remaining cratonic block of the southern part of the Ventuari-Tapajós Province (VTP), a continuation of the VTP to the south of the Alto Tapajós basin (Rizzotto et al., 2019).

The geological framework of the eastern AFMP is separated into five main lithostratigraphic units and has been simplified from Assis et al., 2017 according to new proposals by Rizzotto et al., 2019, into (Fig. 1): (i) 2.86-1.78 Ga Peixoto de Azevedo Domain ; (ii) 1.82-1.78 Ga Colíder Group; (iii) 1.78-1.75 Ga Teles Pires suite; (iv) unconstrained rocks; and (v) 1.98-1.37 Ga Caibis Group, and Phanerozoic sedimentary cover sequences.

Gold deposits in the AFMP are mostly concentrated along with a set of WNW-ESE shear zones (called Peru-Trairão) and correspond to Cenozoic placers, eluvial deposits, and hydrothermal occurrences (Paes de Barros, 2007; Miguel Jr, 2011). From 1980 to 1999, the region produced over 160 t of gold, mostly related to secondary ore sources. Currently, there is a focus on primary types carrying a new exploration cycle in the region and leading to significant discoveries such as Luizão, Pé Quente, Edu, and X1 deposits.

Rocks in the province are mostly represented by Paleoproterozoic oxidized calc-alkaline, medium-to high-K, peraluminous to metaluminous I-type granites, volcanic and volcano-sedimentary sequences, with the presence of subordinate later A-type granites and volcanic rocks. Juliani et al., 2021, grouped the units from AFMP into: i) deformed granitic basement (2.81-1.99 Ga); ii) felsic I-type granites, volcanic and volcano-sedimentary units (1.97-1.87 Ga); iii) highly evolved I-type post-orogenic and A-type and volcanic, sub-volcanic, and volcano-sedimentary sequences; iv) and clastic sedimentary sequences. Spatial analysis indicates a close relationship between the primary gold occurrences and I-type granites (calc-alkaline and oxidized), volcanic rocks, and volcano-sedimentary sequences originated in continental magmatic arcs.

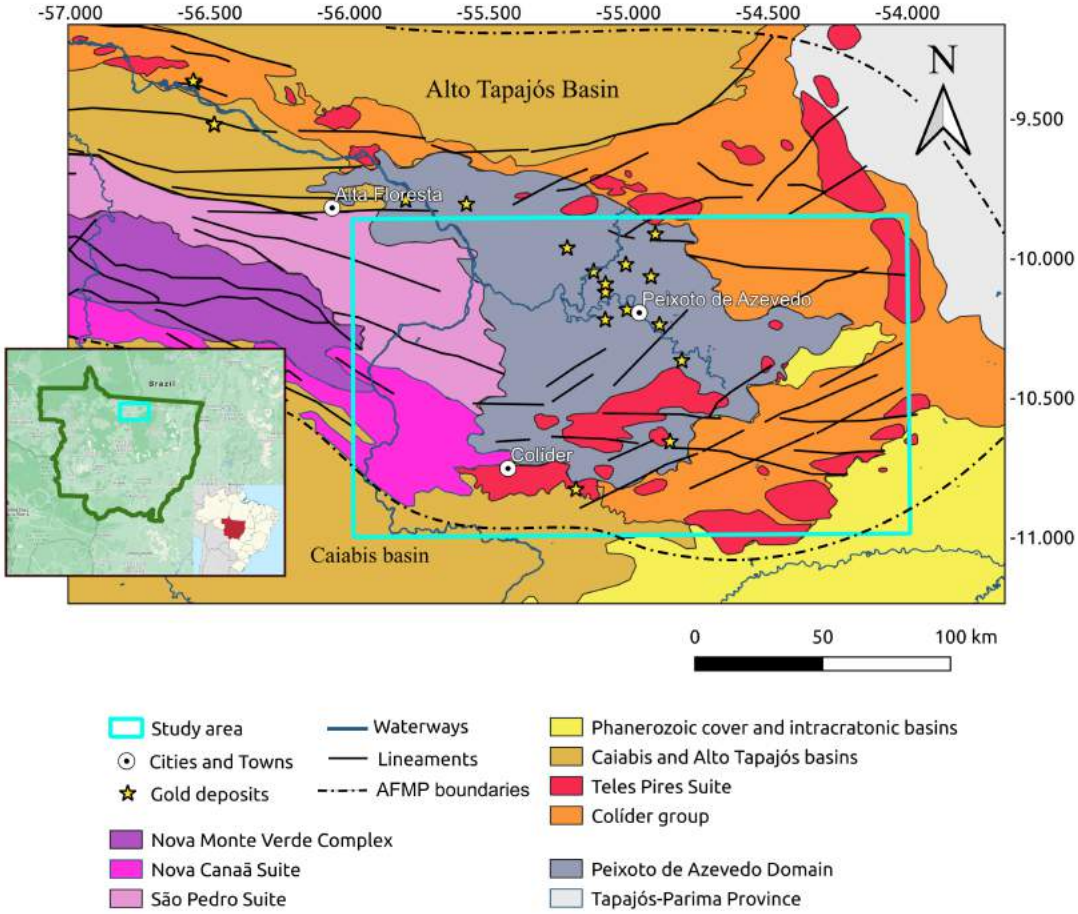


Figure 1: Simplified geological map of the Eastern Alta Floresta Mineral Province, modified from Rizzotto et al., 2019.

The ensemble of hydrothermal alteration, styles, ore paragenesis, fluid inclusion, and isotopic data, suggest the formation of the deposits in Paleoproterozoic magmatic-hydrothermal systems distributed in different crustal levels and positions (Juliani et al., 2021).

The primary gold deposits were summarized by Assis, 2015, based on its tectonic framework, stratigraphy, and genetic models:

(i) Orogenic deposits in metamorphosed magmatic arc rocks in shear zones (e.g., Complex Cuiú–Cuiú) (Barros, 1994; Teixeira, 2015).

(ii) Epizonal deposits related to intrusive felsic rocks (reduced to oxidized)(intrusion-related gold systems – IRGS), disseminated-style or stockwork-style (Abreu, 2004; Paes de Barros, 2007; Bizotto, 2004).

(iii) Porphyry-type deposits, gold mineralizations (disseminated-style) in Paleoproterozoic calc-alkaline granites (Moura et al., 2006; Assis, 2011; Trevisan, 2015).

(iv) Epithermal mineralizations from disseminated-style to confined to venules and veins or stockwork-style, both associated to Paleoproterozoic plutono-volcanic (calc-alkaline to alkaline) (Assis, 2011; Trevisan, 2015).

3 Data and Feature Engineering

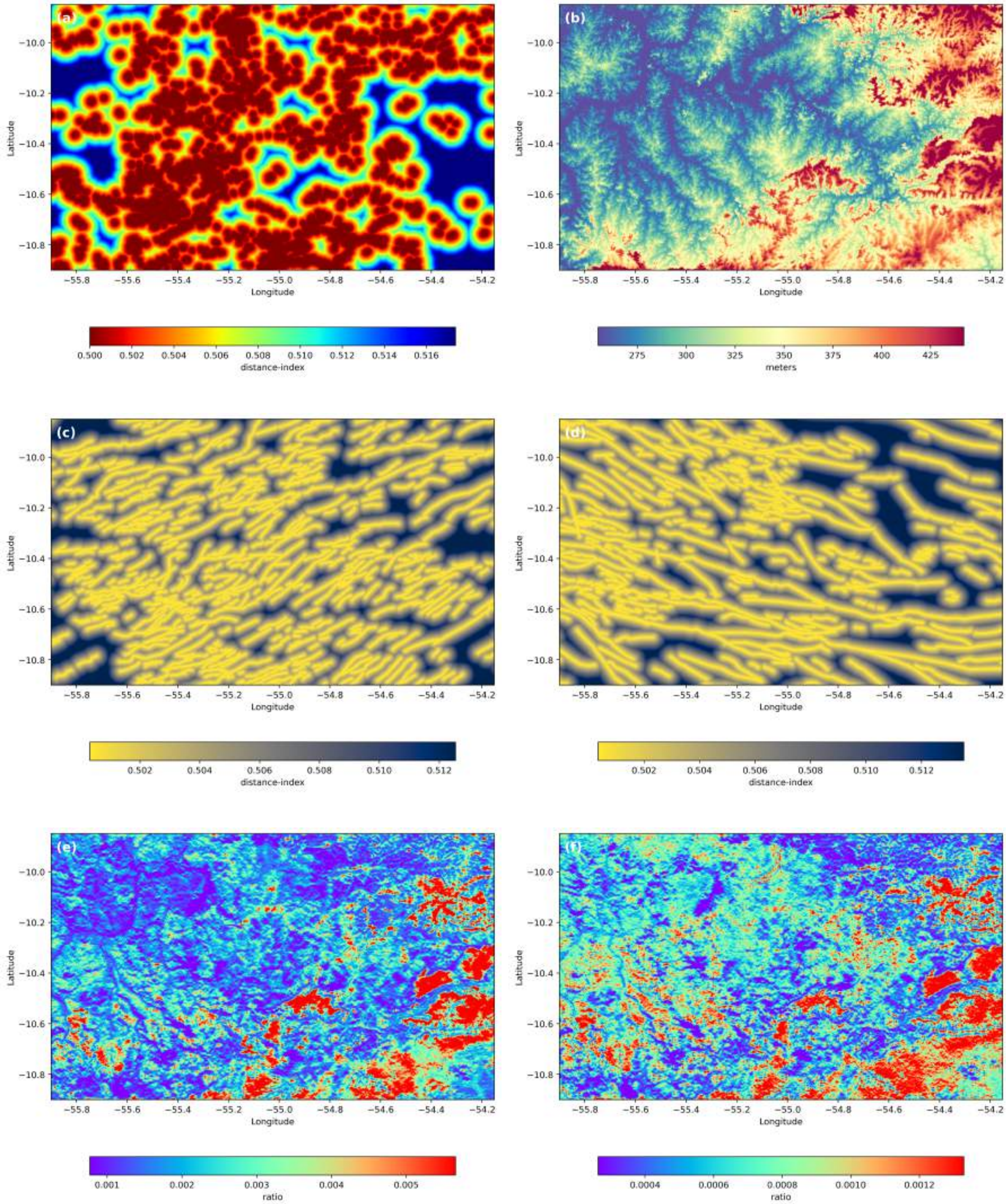
We used open datasets available in the GeoSGB (Geological Survey of Brazil) and Earth Explorer (United States Geological Survey) platforms to generate predictor maps. The feature selection and engineering followed the geological understanding of the gold mineral systems present in AFMP. Diverse exploration data such as airborne geophysics, elevation, geochemistry, and geology, were used to represent the source, transport, alteration and mineral association, which characterize the ore-forming processes in the province.

The airborne magnetic and gamma-ray data belong to the Project 1121 - Norte do Mato Grosso, coordinated by the Geological Survey of Brazil (CPRM) from 2012 to 2014. The survey consists in north-south lines spaced by 500 m, tie-lines separated by 10000 m, terrain clearance of 100 m and data collected each 8 m (magnetic field) and 80 m (gamma-ray spectrometry), approximately. Lateral limiting techniques for magnetic field data, as the 3D Analytic Signal Reid et al., 1990 and the Tilt-Derivative Miller and Singh, 1994, were computed using Fast-Fourier Transform (FFT) in Fatiando a Terra package Uieda et al., 2013. These products allowed the generation of the map of Distance to Magnetic Gradient Maxima (Fig. 1a), and of two maps of Distance to Magnetic Lineaments (Fig. 1c; Fig. 1d). They may correspond to intrusive rock bodies related to the Colider and Teles Pires suites, contemporary to part of the gold mineralizations in AFMP Assis et al., 2017, and candidates to ore transport paths from the source to the deposition sites, respectively.

The gamma-ray data were used to compute the ratios between eTh/K (Fig. 1e) and eU/K (Fig. 1f). This procedure allows highlighting possible potassic alteration or sericitized sites, commonly associated to the gold mineralizations in AFMP Abreu, 2004; Assis, 2015. Thus, the ML algorithms are expected to search for low ratio values to due the differential mobility between these three elements (K $>$ Th and K $>$ U) (Shives et al., 2000).

The digital elevation model (Fig. 1b) was derived from the Shuttle Radar Topographic Mission (SRTM) data, originally with 30 meters of pixel size. The geochemical data from stream sediment were collected during the Mineral Resources project by the Brazilian Geological Survey (CPRM), and present a sampling density from 15 km^2 to 5 km^2. The predictor map of Au-Cu association (Fig. 1g) represents a Geochemical Mineralization Probability Index (GMPI), and was derived from the original geochemical data using the methodology proposed by Yousefi et al., 2012. The sediment catchment areas were delimited using QGIS, and all samples were interpolated according to the corresponding influence regions using a inverse distance weighted method.

All the variables were re-interpolated to an identical regular grid of 200 meters cell size. A Leave-Ball-Out validation (Le Rest et al., 2014) selected the best method among Linear, Cubic, and Nearest Neighbors. The magnetic and K used the cubic method, whereas the eTh and eU data had the Nearest Neighbors as the best method.



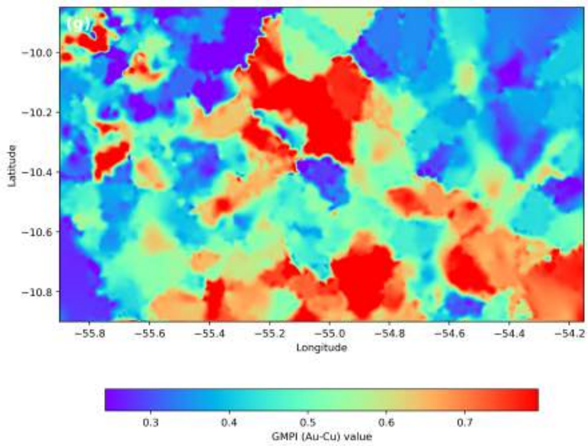


Figure 1: Predictor maps: a) proximity to magnetic gradient maxima; b) digital elevation model (DEM); c) proximity to NE-SW magnetic lineaments and structures; d) proximity to WNW-ESE magnetic lineaments and structures; e) gamma-ray thorium-potassium ratio (Th/K); f) gamma-ray uranium-potassium ratio (U/K); g) Geochemical Mineralization Probability Index associated with gold and copper (Au-Cu).

4 Machine learning methods

4.1 Random forest

Random forest (RF) consists of a combination of tree predictors that depend on values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). It is an ensemble machine learning method that combines multiple decision trees, realizing repeated predictions of the same phenomenon represented by the training dataset (Carranza, 2015). In ensemble methods, for the kth tree, a random vector Θ_k is generated, independently of the past random vectors Θ_1, ..., Θ_{k-1}, but presenting the same distribution. A tree grows using the training set and Θ_k, resulting in a classifier $h(x, \Theta_k)$ where x is an input vector. In dichotomous problem, as the classifications in MPM, the model iteratively splits the dependent variable (root mode) into binary pieces (leaf nodes) in every decision tree. The trees then search along all the splits in order to find the best one that maximizes the purity of the resulting trees (Rodriguez-Galiano et al., 2014).

There are many approximations for measuring impurity. The most frequent used ones are entropy, gain-ratio, Gini index, and chi-square. The entropy, for example, attempts to maximize the mutual information, constructing a equal probability node in the decision (Eq.1):

$$Entropy = -\sum_j p_j \log_2 p_j; \tag{1}$$

While the classification error can be measured as (Eq. 2):

$$ClassificationError = 1 - max\ p_j; \tag{2}$$

where p_j is the probability of class j.

4.2 Gradient boosting

Gradient boosting is a technique for regression and classification problems that produces a prediction model in the form of an ensemble of weak prediction models. A boosted regression or classification uses the gradient descent approach to optimize model parameters. In this approach, the parameters of a model are iteratively moved in the direction of the steepest descent of the loss function. A predictor-corrector uses a chain of models seeking to correct the last model by approximating it to the training set iteratively. It uses weights based on the accuracy of the previous model to increase the probability of obtaining a higher prediction accuracy of the subsequent model. This approach can handle feature interactions well and is flexible in the choice of the loss function (Nwaila et al., 2020).

Friedman, 2001 defines a function estimation problem with a random output y, and a set of explanatory variables $x = x_1, x_2, \ldots, x_n$. Using a training set $\{y_i, x_i\}_1^n$ of known (y, x)-values, the goal is to obtain an estimate or approximation $\hat{F}(x)$ of the function $F^*(x)$ by mapping x to y, which minimizes the expected value of some specified loss function $L(y, F(x))$ over the joint distribution of all (y, x)-values. The loss functions $L(y, F)$ include squared-error $(y-F)^2$ and absolute error $|y-F|$ for $y \in \mathbb{R}^1$ (regression), and negative binomial log-likelihood, $log(1 + e^{-2yF})$, when $y \in -1, 1$ (classification).

4.3 Support Vector Machine

Support vector machines (SVMs) are supervised learning methods used to model datasets in classification and regression problems. The algorithm initially seeks to construct a linear classifier to separate different classes with the widest decision boundaries. In more complex cases, in which the origin feature spaces are not linearly separable, the input data have to be transformed into a higher n-dimensional feature space through various kernel functions, where a classification problem can be performed by a (n-1) dimensional plane referred as hyperplane. A soft margin kind is utilized to minimize the misclassifications (Suykens and Vandewalle, 1999).

For linear problems, in a hyperplane decision function $f(x) = sgn((w \cdot x) + b)$, the Vapnik-Chervonenkis (VC) dimension is controlled by the norm of the weight vector w, and given a training set $(x_1,\ y_1), \ldots, (x_l,\ y_l)$, with $x_i \in \mathbb{R}^n$, and $y_i \in \pm 1$, a border/margin that better separates classes can be found by minimizing $\frac{1}{2}\|w\|^2$ subject to $y_i \cdot ((x_i \cdot w) + b) \geq 1$ for $i = 1, \ldots, l$, (Schölkopf et al., 1998).

In nonlinear problems, for a training dataset defined as $\{x_i\}_{i=1}^n$ with n feature vectors, a labeled target dataset $\{y_i\}_{i=1}^n$ is associated with each vector x_i. The labels, in this study, are $y = 1$, and $y = 0$, indicating respectively, deposits and non-deposits. As the input data cannot be linearly separated in the original feature space, they are mapped into a higher-dimensional space H by a mapping function $\Phi : \mathbb{R}^n \rightarrow H$. (Burges, 1998).

There are four popular kernel functions used in SVM algorithms, which include linear, radial basis, polynomial, and sigmoid function. For geoscientific problems, the radial basis function (RBF) is considered the most suitable due to its simplicity and adaptability (Zuo et al., 2015). The RBF can be described as (Eq. 3):

$$K(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2),\ \gamma > 0, \tag{3}$$

in which γ determines the width of RBF.

4.4 K-Nearest Neighbors

K-nearest neighbors (KNN) is one of the most fundamental and simple classification methods. It aims to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. KNN classifier is commonly based on the Euclidean distance, seeking to segment and find patterns by learning from a training set $(x_{i1}, x_{i2}, \ldots, x_{ip})$, being n the total number of input samples $(i = 1, 2, \ldots, n)$, and p the total number of features $(j = 1, 2, \ldots, p)$(Peterson, 2009). The Euclidean distance between sample x_i and x_l $(l = 1, 2, \ldots, n)$ is defined as in Eq. 4.

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + \cdots + (x_{ip} + x_{lp})^2}. \tag{4}$$

Classification typically involves partitioning samples into training and testing categories. Consider x_i and x, respectively, training and test sample, and let ω be the true class of a training set and $\hat{\omega}$ be the predicted class for a test set $(\omega, \hat{\omega} = 1, 2, \ldots, \Omega)$, where Ω is the total number of classes. During the training process, only the true class ω is used. When testing, the model predicts $\hat{\omega}$ of each sample. For $K = 1$, the predicted class of test sample x is set equal to the class ω of its nearest neighbor, where m_i is a nearest neighbor to x if the Eq. 5 is true.

$$d(m_i, x) = min_j\{d(m_j, x)\}. \tag{5}$$

Generalizing $\forall\ K \in \mathbb{Z}_+^*$, the predicted class of test sample x is set equal to the most frequent true class among K nearest training samples. This forms the decision rule $D :\ x \rightarrow \hat{\omega}$ (Peterson, 2009).

5 Methodology

5.1 Models training

In geoscientific problems, model validation presents issues due to spatial autocorrelation. That happens because the data violate the assumption of independence of most standard statistical procedures, once natural phenomena create spatial continuity of physical/chemical properties (Griffith, 1987; Legendre, 1993). Given this, to avoid over-optimistic results, it is preferable to split the data into training and test sets according to geographic regions. In this paper, the training and test sets are composed only of samples within the north and south portions of the study area, respectively. Point Pattern Analysis (PPA) was performed to ensure that the negative examples could be randomly sampled outside the boundaries with the most likely locations to find new deposits (Fig. 2).

In this study, we used an interval for negative random sampling. In order to simplify, it will be called Negative Random Sampling Interval (NRSI) henceforth. The NRSI consists of an array of distances within the range of least favorable sites to find new deposits, defined through PPA, and must have a size equal to the number of realizations. In this case, it starts at 7.45 km, corresponding to a region with a probability of 3.97% to find new deposits, and ends at 30.8 km (the maximum distance between known deposits in the eastern sector of AFMP). At each iteration, a new distance in NRSI is considered (from smallest to largest) to delimit both likely and unlikely deposit regions. Then, new random negative examples are sampled from smaller areas and less likely are to find deposits. The Fig. 3 illustrates how the distances between known deposits and negative samples become larger with each new iteration. The goal is to analyze the impacts it has on the performance of ML models and the contribution of features.

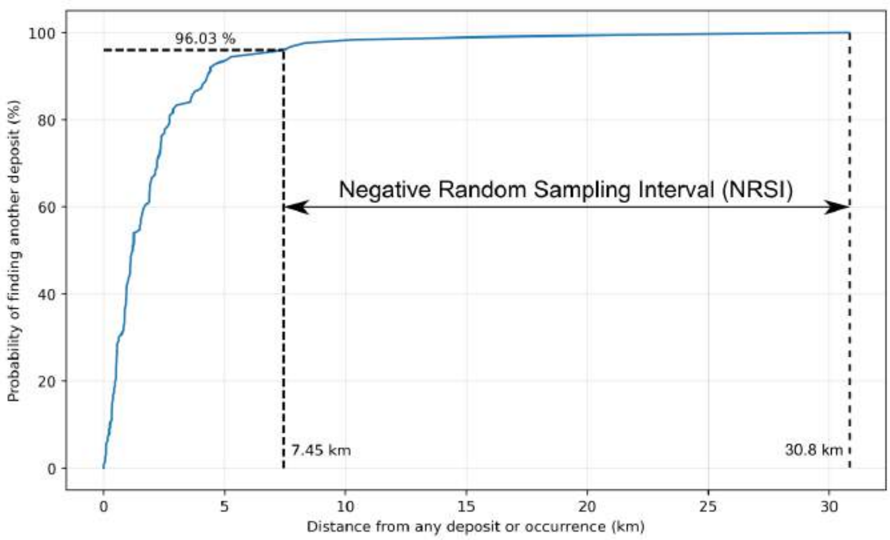


Figure 2: Result of point pattern analysis showing neighboring distance and probability of finding new gold deposits.

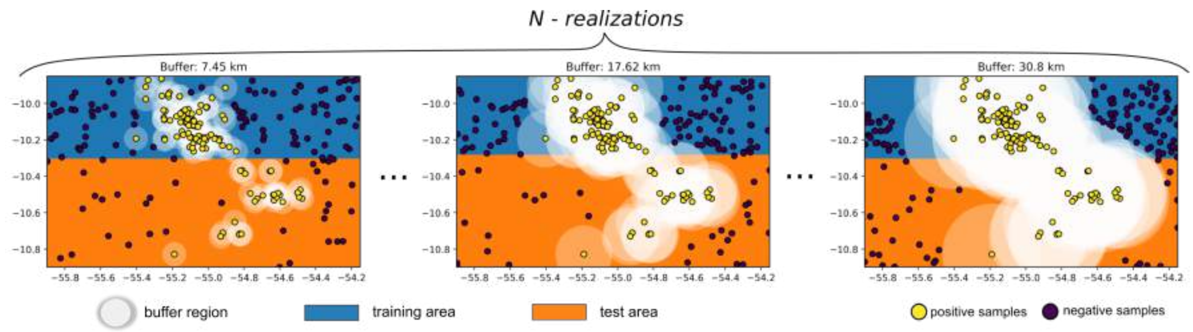


Figure 3: Scheme representing three arbitrary realizations with positive samples (yellow) and negative random samples (dark purple) splitted into two geographic regions: training area (blue) and test area (orange); separated by a increasing deposit location buffers (white).

The training and test sets have 196 samples (98 positive/98 negative) and 58 samples (29 positive/29 negative), respectively. Positive samples include primary gold mineral deposits and occurrences. The optimum hyperparameters (Tab. 1) were obtained using a random search (Bergstra and Bengio, 2012), where the best fit among 250 iterations for 6 folds, totaling 1500 fits, was selected for each type of algorithm.

There are several well-known metrics used to evaluate machine learning classifications, such as accuracy, f1-score, recall, and confusion matrix. In a confusion matrix, for example, four outcomes of a classification are summarized as: (i) a deposit sample correctly classified as a deposit is referred to as true positive sample (TP); (ii) a deposit sample incorrectly classified as a non-deposit is referred to as false negative sample (FN); (iii) a non-deposit samples incorrectly classified as a deposit is referred to as false positive sample (FP); and (iv) a non-deposit sample correctly classified as a non-deposit is referred to as true negative (TN). Based on the confusion matrix, it is possible to derive the other performance metrics (Eq. 6 to 8).

$$\text{Accuracy-score} = \frac{TP + TN}{TP + TN + FP + FN}; \quad (6)$$

$$\text{Recall-score} = \frac{TP}{TP + FN}; \quad (7)$$

$$\text{f1-score} = 2 \times \left(\frac{precision \times recall}{precision + recall} \right). \quad (8)$$

6 Results and discussion

6.1 Models computation and potential maps

One thousand realizations for each ML algorithm (RF, SVM, GB and KNN) were computed (Fig. 4). Each model, at each location, had two maps generated by calculating the mean and variance of the probabilistic outputs (Fig. 5). The variance maps display stabler predictions in some locations to the detriment of others, revealing different risk levels at each predicted location. Furthermore, to evaluate the general performance of the ML models, we stored the accuracy, recall, and f1-score of all realizations. They made it possible to calculate the cumulative metric values according to increasing deposit location buffers, which showed that the predictions tend to converge after a few hundred realizations (Fig. 6).

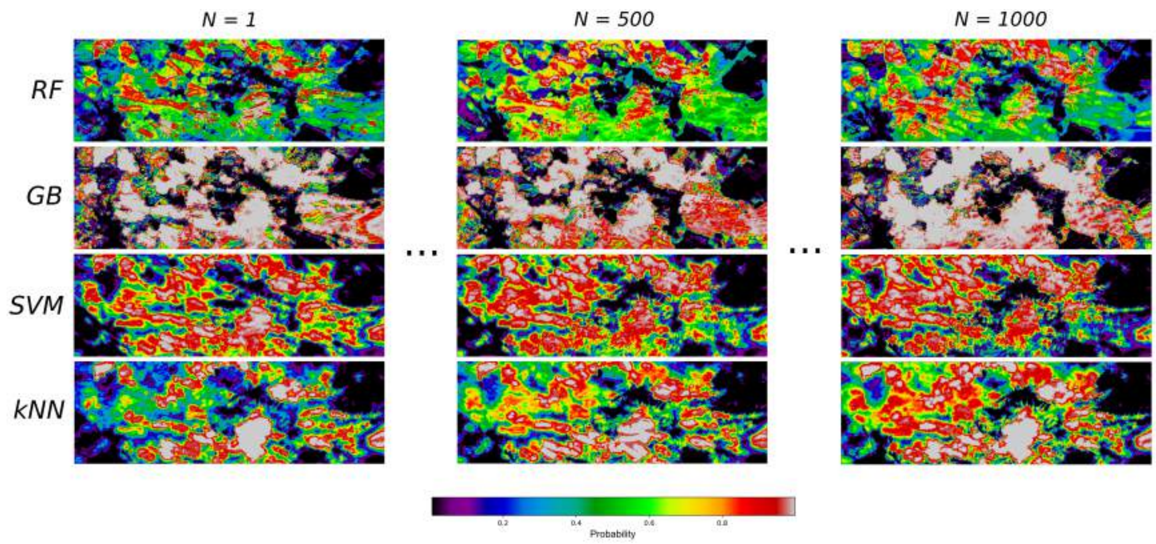


Figure 4: Three arbitrary realizations showing increasing minimum distances to positive samples of the four machine learning models (top to bottom): random forest (RF) predictions; gradient boosting (GB) predictions; support vector machine (SVM); k-nearest neighbors (kNN) predictions.

Table 1: Hyperparameters used for training machine learning models.

RF	n-estimators: 600	min. samples split: 10	min. samples leaf: 4	max. features: *sqrt*	max. depth: 40
	bootstrap: *false*	criterion: *gini*			
GB	n-estimators:100	val. fraction: 0.75	subsample:0.25	min. samples split: 2	min. samples leaf: 2
	max. features: *sqrt*	max. depth: 15	loss: exponential	learning rate: 0.05	criterion: *mae*
SVM	C: 1	γ: 0.1	kernel: *RBF*		
kNN	n-neighbors: 40	p: 2	algorithm: *auto*	weights: *uniform*	

5.2 Models evaluation

As new negative random samples serve as input to the models, and new feature values are used to separate positive from negative classes, the contribution of each feature to the classifications tends to change. Each feature can be separated into two perfectly distinguishable intervals: (i) containing all the positive samples; (ii) containing no positive sample. Comparing two features, A and B, where for A the boundaries with all positive samples contain some negative samples (Fig. 7a) whereas for B the boundaries with all positive samples contain no negative samples (Fig. 7b). Once B presents greater purity of information than A, i.e., the two classes are easier to separate using B, it contributes more than A to the final predictions. We computed the correlations between feature importances and distances in NRSI to analyze such behavior. The

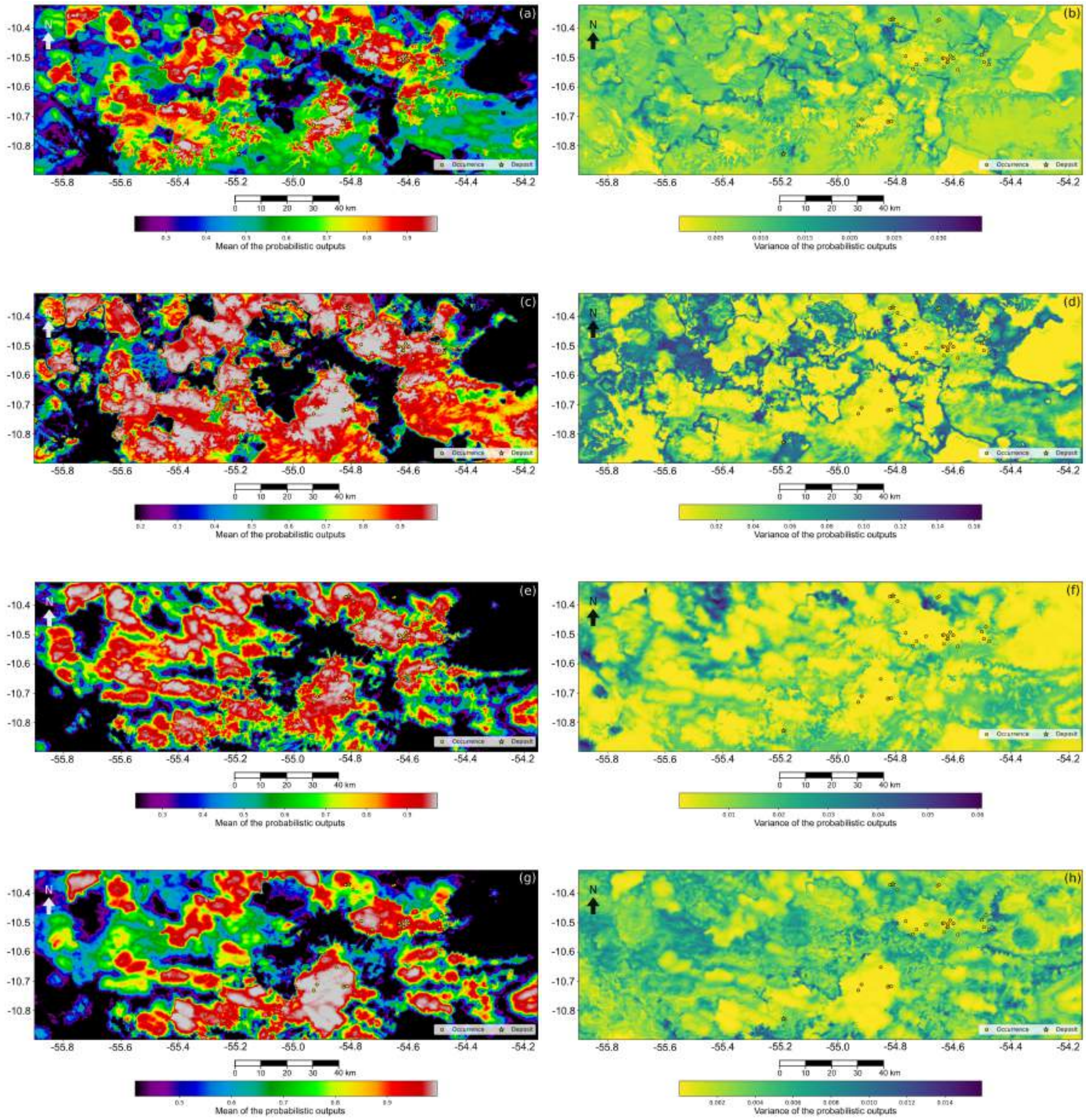


Figure 5: One thousand machine learning model realizations: the mean (a) and variance (b) of RF model predictions; the mean (c) and variance (d) of GB model predictions; the mean (e) and variance (f) of SVM model predictions; the mean (g) and variance (h) of KNN model predictions.

feature importances were derived from shapley values based on their average impact to the final predictions. Seven scatter plots with contribution of features vs distances in NRSI were generated for each ML algorithm (Fig. 8). The correlations suggest that the contribution of the features depend on the buffer of deposit locations. The results indicate that prospectivity mapping is significantly impacted by the scale of the study (further discussed in subsection 6.2).

To understand how the different scenarios impacted the general performance of the ML models, we calculated boxplots of areas under the curve (AUC) to every one hundred realizations (Fig. 9). The tree-based models show stable overall performance, with occasional very high or very low AUC-scores. However, SVM and KNN models

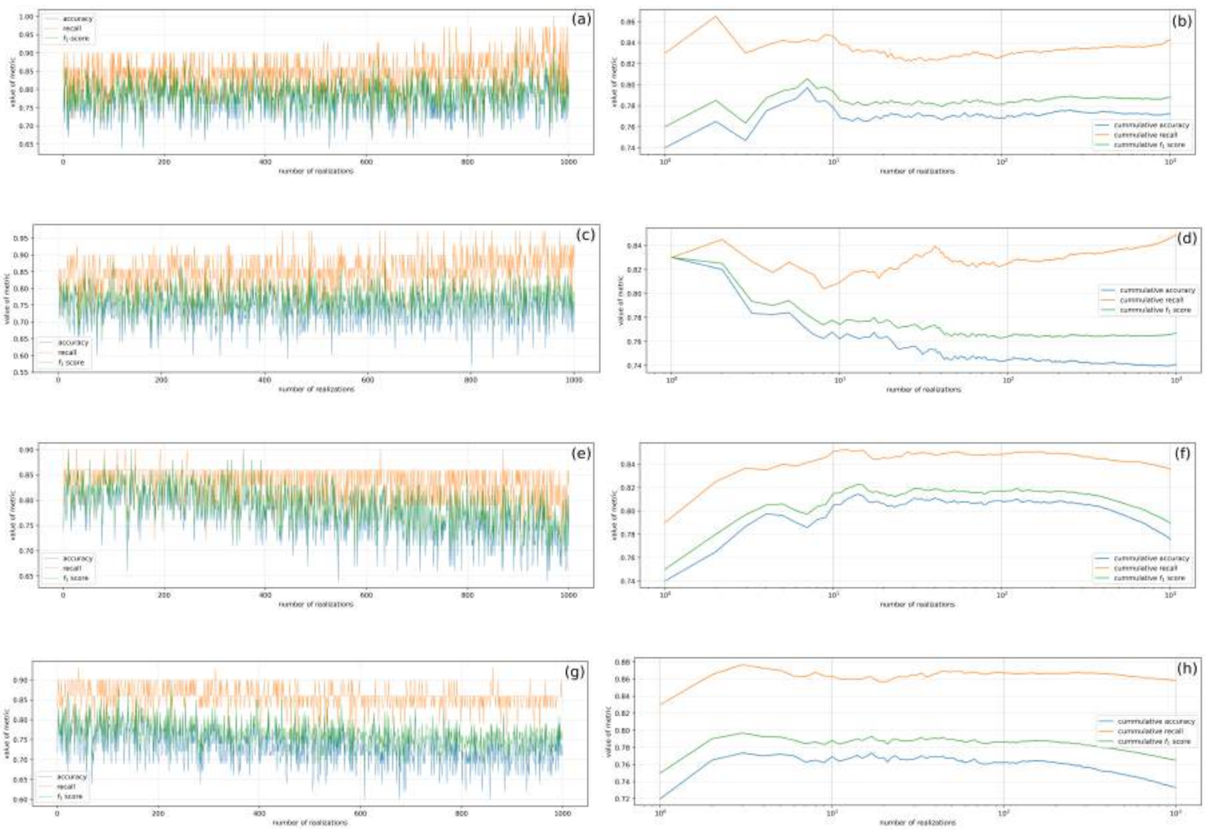


Figure 6: Graphics showing performance metrics (accuracy, recall, and f1-score) of the four machine learning models at each realization: performance metrics per realization (a) and cumulative average value of metrics (b), random forest; performance metrics per realization (c) and cumulative average value of performance metrics (d), gradient boosting; performance metrics per realization (e) and cumulative average value of performance metrics (f), support vector machine; performance metrics per realization (g) and cumulative average value of performance metrics (h), k-nearest neighbor.

display decreasing performances for minimum distances to positive samples farther than 12 km, approximately.

6.2 Prospectivity models *vs.* geological validation

Effective prospectivity maps need models with reasonable input, i.e., proxies representing part of/all the responsible mechanisms for ore formation. We generated seven predictor maps to represent spatial proxies of source, transport, and deposition processes associated with the gold formation in AFMP. Moreover, as important as providing appropriate input information to models, it is crucial to obtain reliable and interpretable outputs. Although, it can be a challenging task due to the "black-box" nature of machine learning modeling processes. Analyzing the contribution of each feature to the final prediction can minimize a poor comprehension of the decisions made by ML models. For this purpose, we computed kernel density functions They display that, in general, the Au-Cu geochemical association has the highest role in the final predictions of all ML models (18 to 55%). However, it also presents a high level of uncertainty.

The elevation model (8 to 30%) had more impact on the district to regional scales

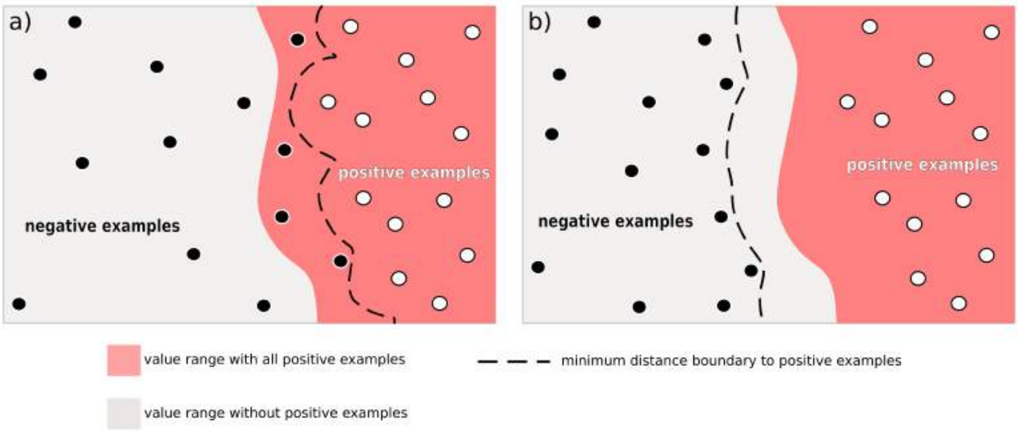


Figure 7: Scheme showing two distinct regions defined by two intervals: one containing all positive examples and another containing no positive examples; a boundary separates negative from positive samples.

(Fig. 8[a2 to d2]). A probable cause for this behavior was the selection of negative samples in locations corresponding to the Cachimbo and Caiabis grabens (elevated areas with low auriferous potential). Lineaments present intermediate contributions (4 to 23%), and those with WNW-ESE orientation are less impacted by the selection of negative samples further from known deposits and mineral occurrences. Besides, they have more impact in deposit to district scales (Fig. 8[a3 to d3]; Fig. 8[a4 to d4]), supporting that structures control from localized gold systems to the occurrence of deposits along with WNW-ESE-striking belts in AFMP.

6.3 Risk-Return analysis

A Risk-Return Analysis framework using ML methods in MPM was proposed by Wang et al., 2020, for the context of assessing uncertainty associated with evidential geological features. Zuo, 2020, adapted such workflow to analyze the effects of using negative random samples in MPM. We applied a similar approach to the predictions of the four ML model, generated throughout multiple realizations, using the methodology presented in subsection 5.1. As in Wang et al., 2020, we captured these uncertainties quantitatively by computing the log-odd ratios of the probabilistic outputs corresponding to the test area. They were compared to the known deposit locations to define four risk-return sectors :11) (i) low risk-low return; (ii) low risk-high return; (iii) high risk-high return; and (iv) high risk-low return. Based on this partitioning, four risk-return maps could be derived (Fig. 12). Comparing the predictions in terms of risk-return, potential maps generated by GB models presented the highest levels of risk and were the most optimistic when defining high return areas (Fig. 11b; Fig.12b). The KNN models displayed low risk levels but definied the smallest high return areas (Fig. 11d; Fig.12d). The predictions of RF (Fig. 11a) and SVM (Fig. 11c) models present intermediate risk levels.

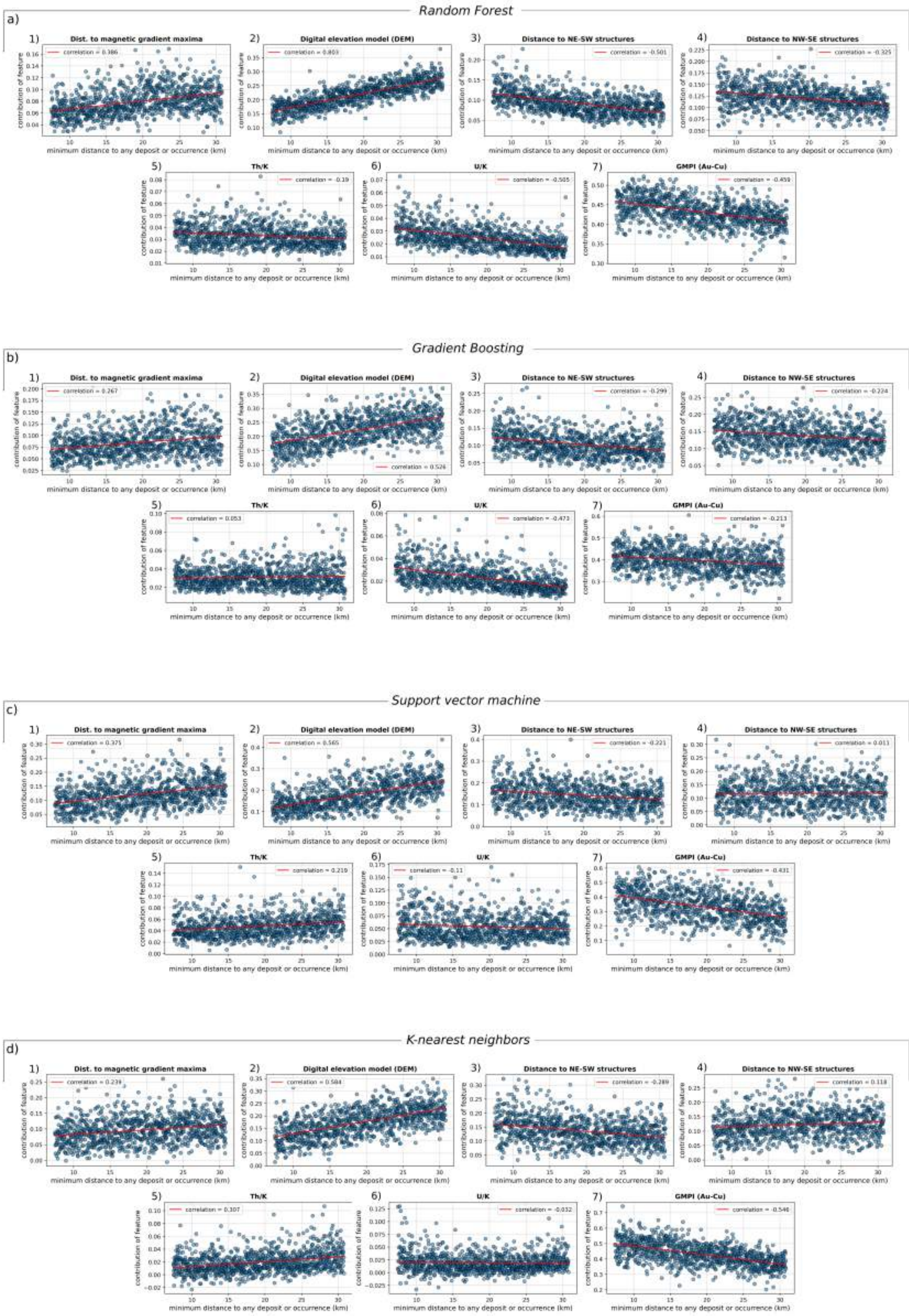


Figure 8: Scatter-plots (feature contribution-minimum distance to any deposit or occurrence) showing fitted lines that represent the trend of contribution of features as the buffer distances increase: a) RF; b) GB; c) SVM; d) KNN.

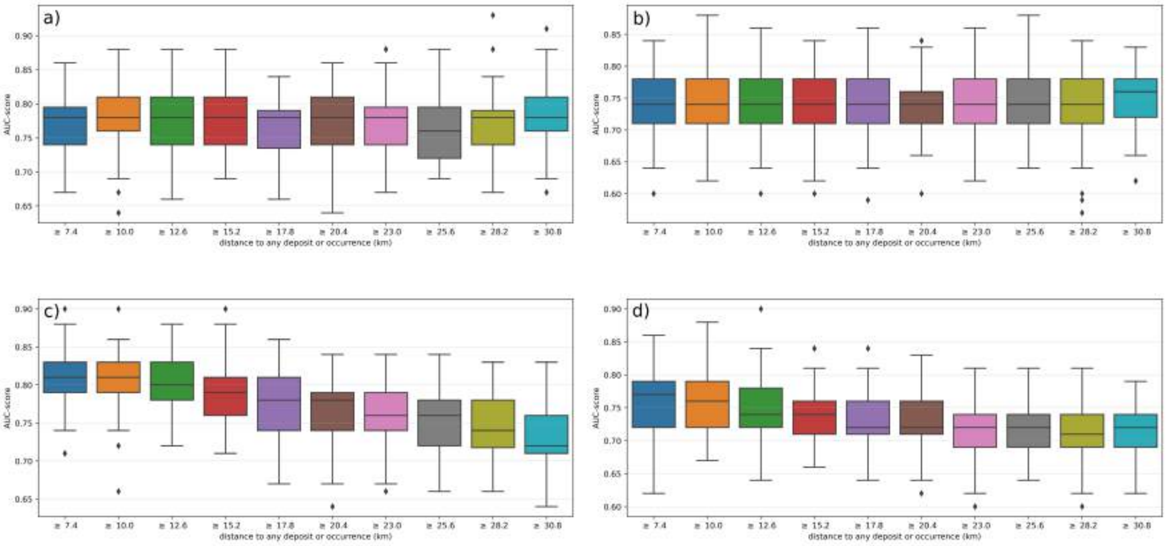


Figure 9: Boxplots representing the area under the curve score (AUC-score) for every one hundred realizations (i.e., 100 realizations per boxplot) of four machine learning models: a) RF; b)B; c) SVM; d) KNN.

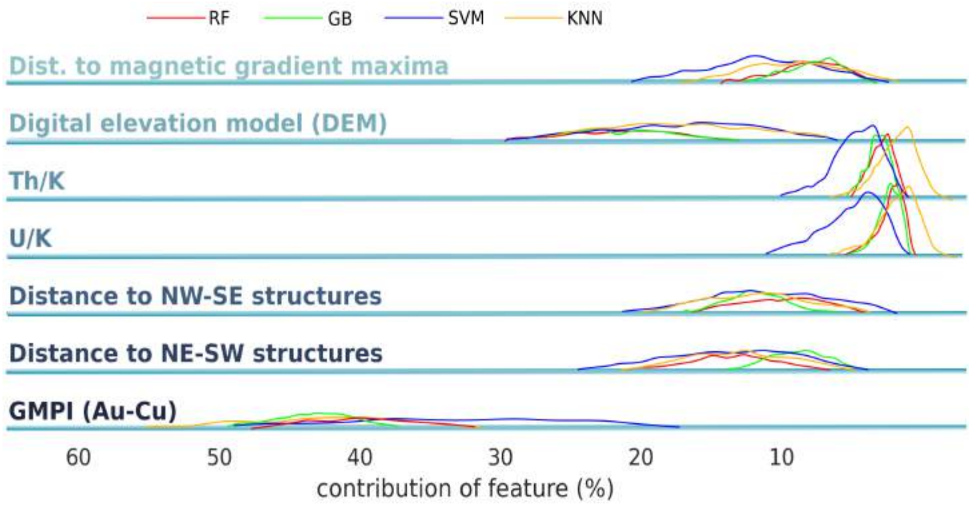


Figure 10: Kernel density functions of the contribution of the features.

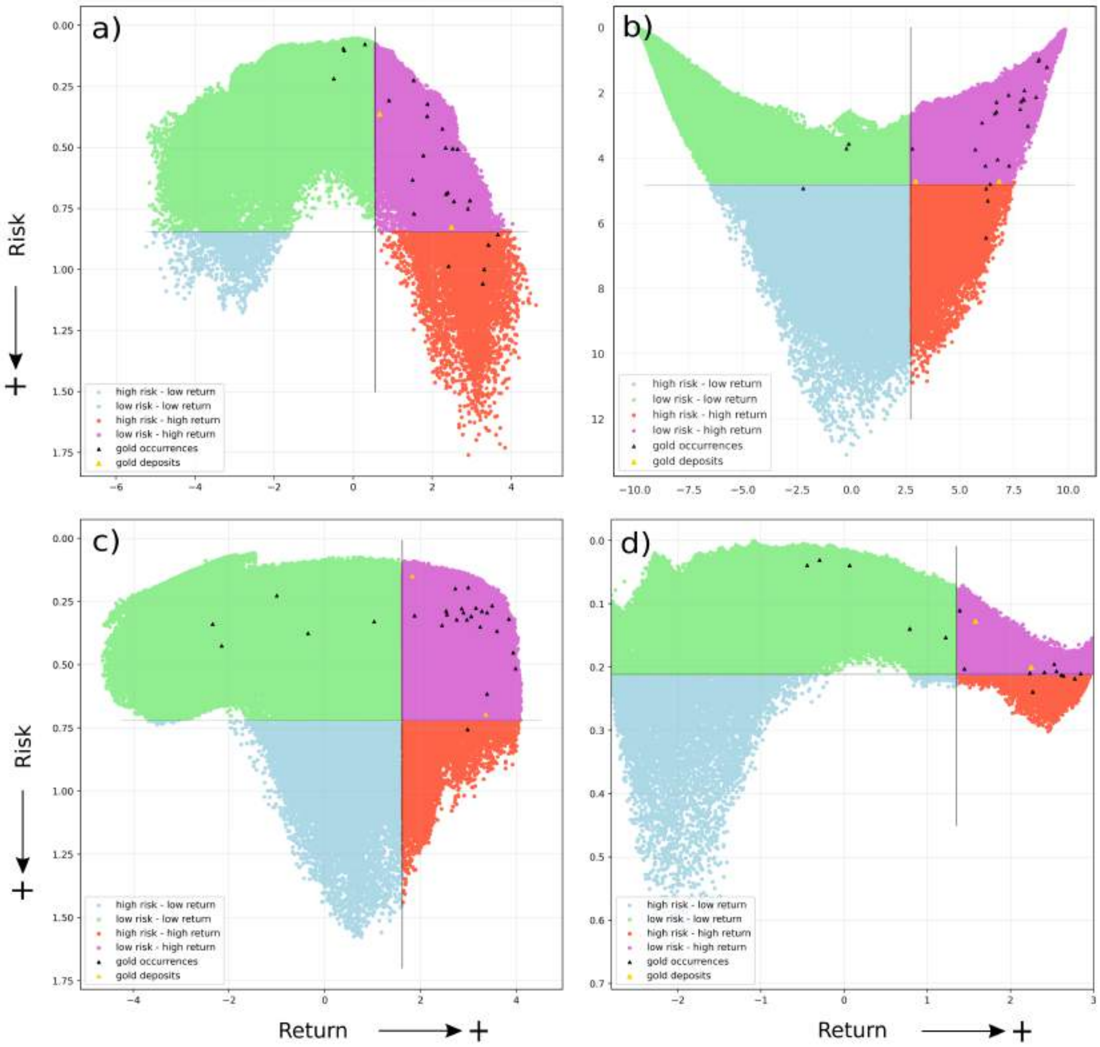


Figure 11: Scatter-plots risk vs. return of four ML models: a) random forest; b) gradient boosting; c) support vector machine; d) k-nearest neighbors. Four risk vs. return sectors: low risk-low return (green); low risk-high return (purple); high risk-low return (blue); high risk-high return (red).

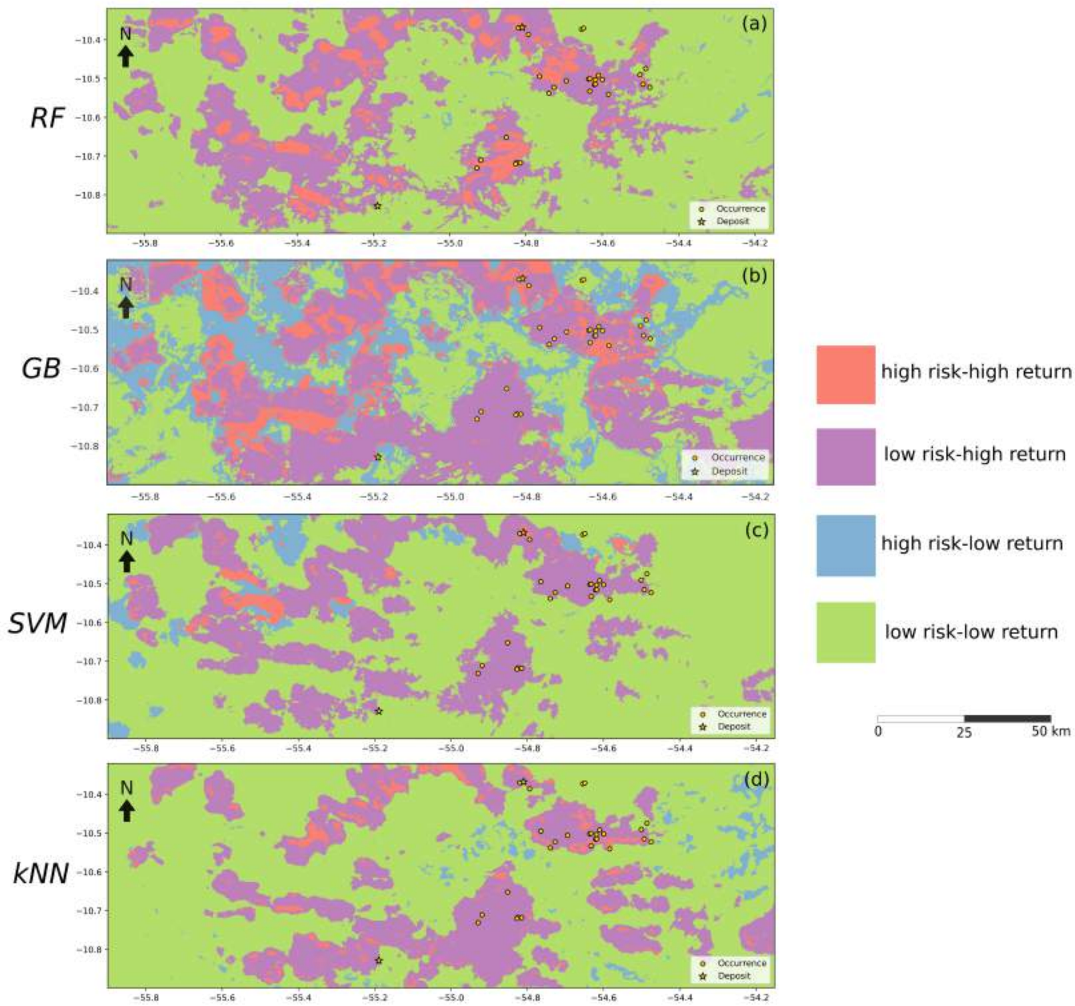


Figure 12: Risk vs. Return maps: a) random forest (RF); b) gradient boosting (GB); c) support vector machine (SVM); d) k-nearest neighbors; separated into four risk-return domains - low risk-low return (green), low risk-high return (purple), high risk-low return (blue), high risk-high return (red).

7 Conclusions

We conducted a risk-return analysis of mineral prospectivity mapping using machine learning methods. The Alta Floresta Mineral Province, Mato Grosso State, Brazil, was chosen as a case study for this research. For uncertainty assessment associated with negative random sampling and buffer of mineral deposits, we computed 1000 realizations of four ML models: random forest, gradient boosting, support vector machine, and k-nearest neighbors algorithms. We evaluated the effects of setting deposit location buffers by selecting negative samples at increasing minimum distances to known mineral deposits and occurrences. All predictions converged after a few hundred realizations. However, the SVM and KNN models presented instability for minimum distances to positive samples greater than 12 km. The contribution of all features displayed significant linear correlations (positives or negatives) with the Negative Random Sampling Interval (NRSI). Four potential maps were generated with the mean and variance of the probabilistic outputs and risk-return analyses were performed using log-odds ratios. The results had four investment sectors defined as low risk-low return, low risk-high return, high risk-low return, and high risk-high return. With such outcomes, we produced four final maps composed of areas with distinct risk-return levels.

Acknowledgments

We would like to thank the Agency of Innovation of the University of São Paulo (AUSPIN) and Brazilian National Council for Scientific and Technological Development (CNPq), that funded this project.

6 Considerações finais

Este projeto nos permitiu revisar as principais aplicações de *machine learning* (ML) no desenvolvimento de mapas de potencial mineral e risco-retorno na Província Mineral de Alta Floresta, possibilitando avanços na explicabilidade de modelos de ML e na análise de incertezas associadas. Em seu desenvolvimento, foram abordadas diversas técnicas de tratamento e integração de dados geofísicos, geológicos e geoquímicos, como Análise de Fatores, Transformada de Fourier, métodos de interpolação e análise de dados espaciais, estatística descritiva entre outras. Dados gamaespectrométricos e magnetométricos, modelo digital de elevação e geoquímica de sedimento de corrente foram utilizados na criação dos vetores de exploração que representem fonte, transporte e alterações relacionadas aos depósitos primários da PMAF. Mapas gradiente de magnetismo máximo, distância até lineamentos magnéticos, elevação, razões Th/K e U/K e associações de Au-Cu foram utilizados na modelagem de caractísticas/assinaturas dos depósitos da província. Mil modelos de ML para simulação de cenários com diferentes favorabilidades auríferas foram gerados utilizando os algoritmos de aprendizado supervisionado RF, GB, SVM e KNN. O cálculo da média e variância das predições realizadas nos permitiu visualizar áreas em que os modelos apresentaram maiorou menor estabilidade e interpretar os resultados obtidos com base nos modelo metalogenético da provincia. As regiões com presença de depósito e/ou ocorrências minerais conhecidas foram classificadas como favoráveis à presença de depósitos de ouro, no entanto os algoritmos GB e SVM apresentaram comportamento super-otimista. Mapas de risco-retorno foram gerados a partir da razão logarítmica de probabilidades e podem ser utilizados na seleção de áreas para investimento e estudos de detalhe. Dentre os vetores de exploração criados, à associação geoquímica de Au-Cu foi dado o maior peso em modelos gerados usando os quatro diferentes algoritmos. O comportamento do peso dados aos vetores de exploração (*features*) ao longo dos cenários simulados sugere relativamente alta correlação espacial entre as áreas consideradas de alta potencial aurífero pelos modelos, que tendem a se concentrar em regiões WNW-ESE ou NE-SW. As regiões consideradas na literatura como pouco favoráveis à descoberta de novos depósitos se mostraram adequadas para a seleção de amostras negativas.

As ferramentas de *machine learning* exibiram grande potencial em aplicações em exploração mineral. Os mapas de potencial mineral se mostraram coerentes quando confrontados com localidades de depósitos de ouro conhecidos. As técnicas de interpretação e o aumento da explicabilidade das predições geradas pelos modelos de ML podem representar ferramentas de auxílio na compreensão do comportamento de depósitos minerais em regiões ainda pouco exploradas. Novas áreas em ascensão, como o uso do *deep learning* para integração de dados, extração de características, geração de modelos 3D, inserção da componente espacial nos modelos etc, têm se apresentado como promissoras e poderiam ser incorporadas em projetos futuros.

7 Bibliografia

Abreu, M., 2004. Alteração hidrotermal e mineralização aurífera do depósito de novo mundo, região de teles pires-peixoto de azevedo, província de alta floresta (mt). Trabalho de Conclusão de Curso, Instituto de Geociências, Universidade Estadual de Campinas, Campinas, 29p.

Almeida, F.F.M., Hasui, Y., Brito Neves, B.B., Fuck, R.A., 1981. Brazilian structural provinces: an introduction. Earth-Science Reviews 17 (1-2), 1–29.

Assis, R.R., 2011. Depósitos auríferos associados ao magmatismo granítico do setor leste da Província de Alta Floresta (MT), Craton Amazônico, tipologia das mineralizações, modelos genéticos e implicações prospectivas. Master's thesis.

Assis, R.R., 2015. Depósitos auríferos associados ao magmatismo félsico da Provincia de Alta Floresta (MT), Craton Amazonico: Litogeoquímica, idade das mineralizações e fonte dos fluidos. Ph.D. thesis, Unicamp.

Assis, R.R., Xavier, R.P., Creaser, R.A., 2017. Linking the Timing of Disseminated Granite-Hosted Gold-Rich Deposits to Paleoproterozoic Felsic Magmatism at Alta Floresta Gold Province, Amazon Craton, Brazil: Insights from Pyrite and Molybdenite Re-Os Geochronology. Economic Geology 112 (8), 1937–1957.

Assis, R.R., Xavier, R.P., de Barros, A.J.P., Barbuena, D., Júnior, E.M., 2012. Contexto geológico e litogeoquímica das unidades plutônicas-vulcânicas da região de União do Norte, setor leste da Província Aurífera de Alta Floresta (MT): Geological setting and geochemistry of the plutonic-volcanic units from the União do Norte region, eastern portion of the Alta Floresta Gold Province (MT). Revista Brasileira de Geociências 42 (1), 130–161.

Barros, A.J.P.d., 1994. Contribuição a geologia e controle das mineralizações auríferas de Peixoto de Azevedo-MT. Instituto de Geociências, Universidade de São Paulo. São Paulo-SP. Master's thesis, Dissertação (Mestrado em Geologia Econômica e Hidrogeologia).

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of machine learning research 13 (2).

Bizotto, A., 2004. Mineralizações auríferas na região de teles pires–peixoto de azevedo (mato grosso), província de alta floresta: O depósito de santa helena. Trabalho de Conclusão de Curso, Instituto de Geociências, Universidade Estadual de Campinas, Campinas, 38p.

Bonham-Carter, G.F., 2014. Geographic information systems for geoscientists: modelling with GIS. Vol. 13. Elsevier.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2 (2), 121–167.

Carranza, Emmanuel John M; Laborte, A.G., 2015. Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm. Ore Geology Reviews 71, 777–787.

Carranza, E., Hale, M., Faassen, C., 2008. Selection of coherent deposit-type locations and their application in data-driven mineral prospectivity mapping. Ore geology reviews 33 (3-4), 536–558.

Carranza, E.J.M., 2009. Controls on mineral deposit occurrence inferred from analysis of their spatial pattern and spatial association with geological features. Ore Geology Reviews 35 (3-4), 383–400.

Caté, A., Perozzi, L., Gloaguen, E., Blouin, M., 2017. Machine learning as a tool for geologists. The Leading Edge 36 (3), 215–219.

Diggle, Peter J; Milne, R.K., 1983. Bivariate Cox processes: some models for bivariate spatial point patterns. Journal of the Royal Statistical Society: Series B (Methodological) 45 (1), 11–21.

Fernandes, C.M., Fontes, S.L., La Terra, E.F., Miquelutti, L.G., Maurya, V.P., 2018. Electrical crustal structure of Alta Floresta Gold Province eastern sector, SW Amazon Craton, Brazil. Journal of South American Earth Sciences 86, 244–258.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232.

Fry, N., 1979. Random point distributions and strain measurement in rocks. Tectonophysics 60 (1-2), 89–105.

Gatrell, A.C., Bailey, T.C., Diggle, P.J., Rowlingson, B.S., 1996. Spatial point pattern analysis and its application in geographical epidemiology. Transactions of the Institute of British geographers, 256–274.

Griffith, D.A., 1987. Spatial autocorrelation. A Primer (Washington, DC, Association of American Geographers).

Carranza, 2008. Geochemical anomaly and mineral prospectivity mapping in GIS. Elsevier.

Juliani, C., Carneiro, C.d.C., Carreiro-Araújo, S.A., Fernandes, C., Monteiro, L., Crósta, A., 2013. Estruturação dos arcos magmáticos paleoproterozoicos na porção sul do Cráton Amazônico: implicações geotectônicas e metalogenéticas. Simpósio de Geologia da Amazônia 13, 157–160.

Juliani, C., Rodrigues de Assis, R., Virgínia Soares Monteiro, L., Marcello Dias Fernandes, C., Eduardo Zimmermann da Silva Martins, J., Ricardo Costa e Costa, J., 2021. Gold in Paleoproterozoic (2.1 to 1.77 Ga) Continental Magmatic Arcs at the Tapajós and Juruena Mineral Provinces (Amazonian Craton, Brazil): A New Frontier for the Exploration of Epithermal–Porphyry and Related Deposits. Minerals 11 (7), 714.

Junior, E., de SantAna Barros, M.A., Pinho, F.E., 2013. Timing of gold metallogeny in the Alta Floresta Gold Province: evidence from pyrite and molybdenite Re-Os isotopic dating.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., Bretagnolle, V., 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Global ecology and biogeography 23 (7), 811–820.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74 (6), 1659–1673.

Mandelbrot, B.B., 1982. The fractal geometry of nature. Vol. 1. WH freeman New York.

Miguel Jr, E., 2011. Mineralizações auríferas do lineamento Peru-Trairão, Província Aurífera de Alta Floresta-MT: controle estrutural e idade U-Pb das rochas hospedeiras. Master's thesis.

Miller, H.G., Singh, V., 1994. Potential field tilt—a new concept for location of potential field sources. Journal of applied Geophysics 32 (2-3), 213–217.

Moura, M.A., Botelho, N.F., Olivo, G.R., Kyser, T.K., 2006. Granite-related Paleoproterozoic, Serrinha gold deposit, Southern Amazonia, Brazil: hydrothermal alteration, fluid inclusion and stable isotope constraints on genesis and evolution. Economic Geology 101 (3), 585–605.

Nwaila, G.T., Zhang, S.E., Frimmel, H.E., Manzi, M.S., Dohm, C., Durrheim, R.J., Burnett, M., Tolmay, L., 2020. Local and Target Exploration of Conglomerate-Hosted Gold Deposits Using Machine Learning Algorithms: A Case Study of the Witwatersrand Gold Ores, South Africa. Natural Resources Research 29 (1), 135–159.

Nykänen, V., Lahti, I., Niiranen, T., Korhonen, K., 2015. Receiver operating characteristics (ROC) as validation tool for prospectivity models—A magmatic Ni–Cu case study from the Central Lapland Greenstone Belt, Northern Finland. Ore Geology Reviews 71, 853–860.

Paes de Barros, Antonio João, e.a., 2007. Granitos da região de Peixoto de Azevedo: Novo Mundo e mineralizações auríferas relacionadas-Província Aurífera Alta Floresta (MT).

Peterson, L.E., 2009. K-nearest neighbor. Scholarpedia 4 (2), 1883.

Porwal, A., Carranza, E., Hale, M., 2003. Artificial neural networks for mineral-potential mapping: a case study from Aravalli Province, Western India. Natural resources research 12 (3), 155–171.

QGIS Development Team, 2009. QGIS Geographic Information System. Open Source Geospatial Foundation.

Reid, A.B., Allsop, J., Granser, H., Millett, A.t., Somerton, I., 1990. Magnetic interpretation in three dimensions using Euler deconvolution. Geophysics 55 (1), 80–91.

Rizzotto, G.J., Alves, C.L., Rios, F.S., Barros, M.A.d.S., 2019. The Western Amazonia Igneous Belt. Journal of South American Earth Sciences 96, 102326.

Rodriguez-Galiano, V., Chica-Olmo, M., Chica-Rivas, M., 2014. Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain. International Journal of Geographical Information Science 28 (7), 1336–1354.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews 71, 804–818.

Santos, J.O.S., 2003. Geotectônica dos escudos das Guianas e Brasil-Central. In: Geologia, tectônica e recursos minerais do Brasil. Vol. 4. CPRM Brasília, pp. 169–226.

Santos, J.O.S., Hartmann, L.A., Faria, M.d., Riker, S.R., Souza, M.d., Almeida, M.E., McNaughton, N.J., 2006. A compartimentação do Cráton Amazonas em províncias: avanços ocorridos no período 2000-2006. Simpósio de geologia da Amazônia 9, 2006.

Santos, J.O.S., Hartmann, L.A., Gaudette, H.E., Groves, D.I., Mcnaughton, N.J., Fletcher, I.R., 2000. A new understanding of the provinces of the Amazon Craton based on integration of field mapping and U-Pb and Sm-Nd geochronology. Gondwana Research 3 (4), 453–488.

Scandolara, J., Correa, R., Fuck, R., Souza, V., Rodrigues, J., Ribeiro, P., Frasca, A., Saboia, A., Lacerda Filho, J., 2017. Paleo-Mesoproterozoic arc-accretion along the southwestern margin of the Amazonian craton: The Juruena accretionary orogen and possible implications for Columbia supercontinent. Journal of South American Earth Sciences 73, 223–247.

Schölkopf, B., Simard, P., Smola, A.J., Vapnik, V., 1998. Prior knowledge in support vector kernels. In: Advances in neural information processing systems. pp. 640–646.

Shives, R.B., Charbonneau, B., Ford, K.L., 2000. The detection of potassic alteration by gamma-ray spectrometry—recognition of alteration related to mineralization. Geophysics 65 (6), 2001–2011.

Silva, F.R.d., Barros, M.A.S., Pierosan, R., Pinho, F.E.C., Rocha, M.L.B.P., Vasconcelos, B.R., Dezula, S.E.M., Tavares, C., Rocha, J., 2014. Geochemistry and geochronology U-Pb Shrimp of granites from Peixoto de Azevedo: Alta Floresta Gold Province. Brazilian Journal of Geology 44 (3), 433–455.

Sun, T., Chen, F., Zhong, L., Liu, W., Wang, Y., 2019. GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China. Ore Geology Reviews 109, 26–49.

Suykens, J.A., Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural processing letters 9 (3), 293–300.

Tassinari, C.C., Macambira, M.J., 1999. Geochronological provinces of the Amazonian Craton. Episodes-Newsmagazine of the International Union of Geological Sciences 22 (3), 174–182.

Tassinari, C.C.G., Macambira, M.J.B., 2004. A evolução tectônica do Craton Amazônico. In: Congresso Brasileiro de Geologia. SBG.

Teixeira, R., 2015. Rochas hospedeiras e controle estrutural da mineralização aurífera do depósito Peteca, região de Flor da Serra–MT, na província Aurífera de Alta Floresta. Trabalho de Conclusão de Curso, Instituto de Geociências, Universidade Estadual de Campinas, 107p.

Trevisan, V.G., 2015. Estudo comparativo entre mineralizações filonares de Au$\pm$Cu e Au+ metais base do setor leste da Província Aurífera de Alta Floresta (MT), Cráton Amazônico. Master's thesis.

Trevisan, V.G., Hagemann, S.G., Loucks, R.R., Xavier, R.P., Motta, J.G., Parra-Avila, L.A., Petersson, A., Gao, J.F., Kemp, A.I., Assis, R.R., 2021. Tectonic switches recorded in a Paleoproterozoic accretionary orogen in the Alta Floresta Mineral Province, southern Amazonian Craton. Precambrian Research 364, 106324.

Uieda, L., Oliveira Jr, V.C., Barbosa, V.C., 2013. Modeling the Earth with fatiando a terra. In: Proceedings of the 12th Python in Science Conference. pp. 96–103.

Van Rossum, Guido e Drake Jr, F.L., 1995. Python Organization. Centrum voor Wiskunde en Informatica Amsterdam.

Wang, Z., Yin, Z., Caers, J., Zuo, R., 2020. A Monte Carlo-based framework for risk-return analysis in mineral prospectivity mapping. Geoscience Frontiers 11 (6), 2297–2308.

Yousefi, M., Kamkar-Rouhani, A., Carranza, E.J.M., 2012. Geochemical mineralization probability index (GMPI): A new approach to generate enhanced stream sediment geochemical evidential map for increasing probability of success in mineral potential mapping. Journal of Geochemical Exploration 115, 24 – 35.

Zuo, R., Carranza, E.J.M., 2011. Support vector machine: a tool for mapping mineral prospectivity. Computers & Geosciences 37 (12), 1967–1975.

Zuo, R., Zhang, Z., Zhang, D., Carranza, E.J.M., Wang, H., 2015. Evaluation of uncertainty in mineral prospectivity mapping due to missing evidence: a case study with skarn-type Fe deposits in Southwestern Fujian Province, China. Ore Geology Reviews 71, 502–515.

Zuo, Renguang; Wang, Z., 2020. Effects of random negative training samples on mineral prospectivity mapping. Natural Resources Research 29 (6), 3443–3455.