# Aprendizado de Máquina

Os modelos de aprendizado de máquina aplicados neste trabalho fizeram uso de dados de espectrometria e gravimetria. DESCREVER DE ONDE VIERAM

A seguir, são apresentados os dados de espectrometria, bem como breve descrição de seu significado e tipo de variável:

* Fator F: parâmetro de Efimov (valores altos para rochas alteradas por fluidos que carregam metais) (float)
* Kd: abundância de potássio normalizado pelo tório (float)
* Ud: abundância de urânio normalizado pelo tório (float)

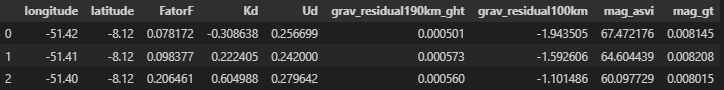
Já os dados de gravimetria, bem como breve descrição e tipo de variável, são apresentados a seguir:

* grav\_residual190km\_qht: gravidade residual a 190 km (float)
* grav\_residual100km: gravidade residual a 100 km (float)
* maq\_asvi: amplitude do sinal analítico da integral vertical do campo magnético (float)
* maq\_qt: gradiente total do campo magnético (float)

A base de dados sem processamento é composta de 116.745 pontos de espectrometria e 114.011 de gravimetria, os quais são expressos em coordenadas UTM (Universal Transverse Mercator), sistema de projeção de mapa para atribuir coordenadas a locais na superfície da Terra. A princípio, a primeira mesclagem da base de dados apresentou redução substancial da base, resultando em apenas um ponto em comum, devido à diferença das grades das bases.

Assim, de modo a não perder uma quantidade significativa de dados, realizou-se uma conversão das coordenadas UTM para latitude e longitude, com um arredondamento de 2 casas decimais, o que trouxe ganhos do ponto de vista de base de dados, porém imprecisão do ponto de vista locacional. Após aplicada tal conversão, a mesclagem da intersecção entre coordenadas dos dados de espectrometria e gravimetria resultou em 113.918 dados. A seguir é apresentada uma captura de tela dos dados resultantes:

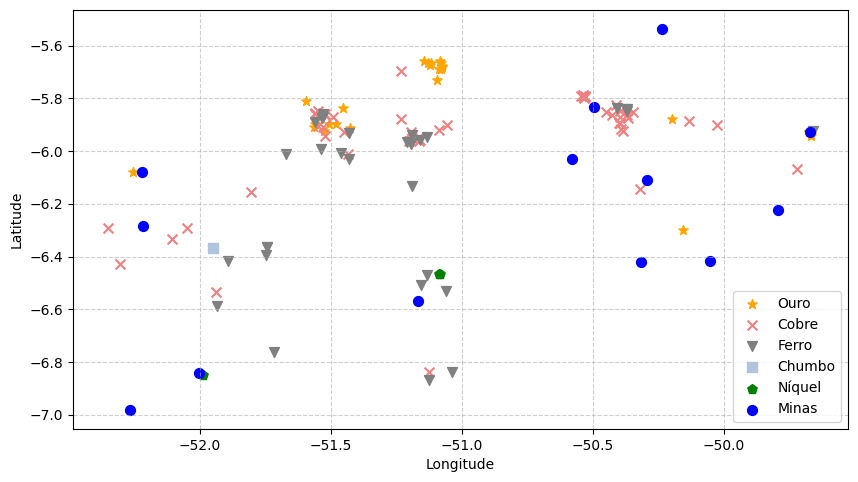
Figura 1 – Base de dados de espectrometria e gravimetria.



Fonte: Própria autoria.

DESCREVER DE ONDE VEM OS MINERAIS. A figura a seguir apresenta um detalhamento de alguns minerais, com suas coordenadas também transformadas de UTM para latitude e longitude.

Figura 2 – Ocorrências de minerais.



Fonte: Própria autoria.

Neste caso, de posse das coordenadas geográficas das ocorrências dos minerais também convertidas para latitude e longitude, bastou realizar outra mesclagem da base de dados anteriormente citada.

Tendo como base a figura de ocorrência de minerais, buscou-se prever a ocorrência de outros possíveis locais com ouro e cobre. Para o primeiro, têm-se 22 ocorrências, e para o segundo, 42 ocorrências.

De pronto, observa-se a permanência de grande desbalanço entre as classes, visto que o a porcentagem de ocorrência para o ouro seria de , resultando em cerca de 0,2% de presença de ouro. De maneira similar, o resultado para o cobre seria de cerca de 0,4%. Neste contexto, o problema seria mais voltado para detecção de anomalias do que problema de aprendizado supervisionado, conforme visto nas disciplinas durante o curso. Assim, alguns procedimentos foram aplicados de modo a reduzir o desbalanceamento entre classes, sendo eles redução da área considerada a priori e reamostragem dos dados. Ressalta-se que em procedimento normal, tais passos não devessem ser realizados, considerando-se uma melhor compatibilidade entre as bases de dados consideradas. Todavia, salienta-se que tais abordagens foram aplicadas devido ao alto desbalanceamento entre dados, o que impossibilitou, a priori, a detecção de minerais, conforme abordagens expressas a seguir, onde serão apresentados todos os pipelines realizados para o estudo de previsão de ouro e cobre, visto que o problema de desbalanceamento é similar a ambos.

O pipeline inicial de aprendizado de máquina se deu conforme a abordagem a seguir:

Figura 3 – Abordagem inicial de pipeline de aprendizado de máquina.

Gráfico, Diagrama

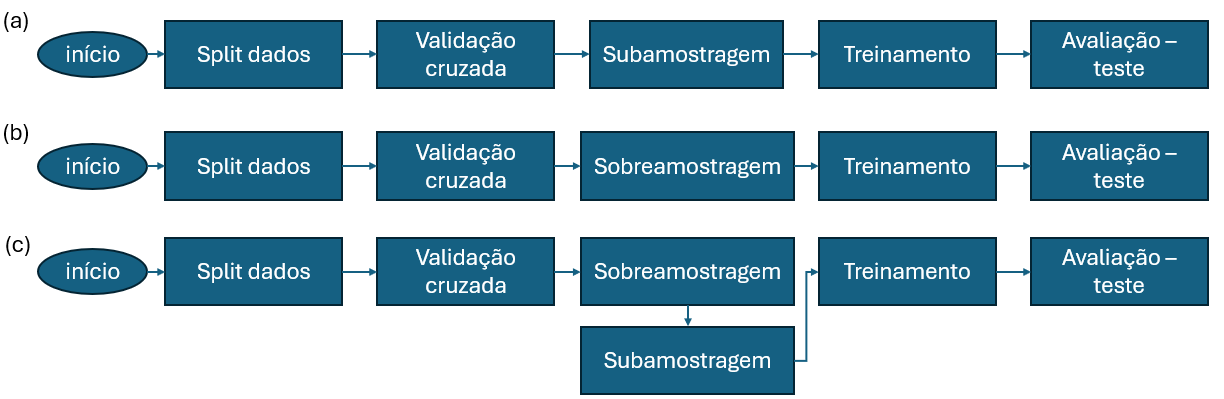
O conteúdo gerado por IA pode estar incorreto.

Fonte: Própria autoria.

A divisão de dados utilizou 70% dos dados para treinamento e 30% para teste, utilizando o parâmetro de estratificação como verdadeiro. A validação cruzada foi feita utilizando a classe StratifiedKFold da biblioteca scikit learn, a qual além de dividir os conjuntos em bases iguais, no caso 5, garante que cada divisão tenha a mesma proporção de dados do dataset original. Tais pontos são importantes, considerando o problema de desbalanceamento.

Tendo em vista os resultados, como será apresentado nos itens a seguir, buscou-se abordagens alternativas de pipeline para melhorar os resultados, aplicando-se (a) subamostragem (*undersampling*) da classe majoritária (não ocorrência de mineral); (b) sobreamostragem (*oversampling*) da classe minoritária (ocorrência de mineral) com a técnica SMOTE (RUTECKI, 2025); e (c) sobreamostragem e subamostragem juntas com as técnicas SMOTE e Tomek Links (RUTECKI, 2025), conforme a figura a seguir:

Figura 4 – Abordagens alternativas de pipeline de aprendizado de máquina.



Fonte: Própria autoria.

Para ambos os estudos, foram utilizados quatro modelos com as seguintes configurações de hiperparâmetros:

* Random Forest com class\_weight como balanced;
* Regressão Logística com class\_weight como balanced, penalidade L2 e solver lbfgs;
* SVM com class\_weight como balanced, kernel rbf, C igual a 1, gamma como scale e probability como verdadeiro;
* XGBoost com scale\_pos\_weight como sendo a proporção entre classes e métrica como logloss.

Todo o processo foi feito em linguagem Python, na IDE VSCode, fazendo uso de bibliotecas padrão como pandas, numpy, scikit learn, matplotlib, seaborn, xgboost, shap e folium, sendo estas duas últimas utilizados para análise das *features* algoritmo de aprendizado de máquina e criação de gráficos iterativos.

## Ocorrências de ouro

A figura a seguir apresenta uma captura de tela do dataframe utilizado para predição de ocorrências de ouro. Observa-se que a coluna de ocorrência é uma variável inteira binária.

Figura 5 – Dados utilizados para previsão de ouro.

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Fonte: Própria autoria.

As figuras a seguir apresentam os passos aplicados para redução do desbalanceamento, conforme mencionado anteriormente. Observa-se que a nuvem de pontos, anteriormente completa na primeira figura, reduz drasticamente na segunda, melhorando assim o balanço entre o número de dados.

Figura 6 – (a) Área inicial e (b) área recortada com reamostragem para redução de desbalanceamento para ocorrências de ouro.

|  |
| --- |
| (a) |
| (b) |

Fonte: Própria autoria.

De maneira a ilustrar a redução, a base de dados final para modelagem é representada pela figura a seguir. Observam-se um total de 717 pontos, sendo 22 deles de presença de ouro, ou seja, 3,07% do dataset, uma grande elevação em relação ao mencionado anteriormente, de 0,2%.

Figura 7 – Contagem dos dados por ocorrência, sendo 1 a ocorrência de ouro.

Gráfico, Gráfico de barras, Histograma

O conteúdo gerado por IA pode estar incorreto.

Fonte: Própria autoria.

Assim, aplicando-se os pipelines comentados na seção anterior, para os modelos Random Forest, Logistic Regression, SVM e XGBoost, os seguintes resultados para predições de ouro (classe minoritária representada como 1 na coluna de ocorrências do dataset) foram obtidas:

Tabela 1 – Métricas dos modelos por pipeline para ocorrências de ouro.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **Abordagem** | **ROC-AUC** | **F1-Score** | **Precision** | **Recall** | **Threshold** | **Acurácia** |
| **Random Forest** | Pipeline Inicial | 0,758 | 0,000 | 0,000 | 0,000 | 0,280 | 0,970 |
| Undersampling | 0,738 | 0,080 | 0,040 | 0,570 | 0,770 | 0,570 |
| SMOTE | 0,697 | 0,267 | 0,250 | 0,290 | 0,460 | 0,950 |
| SMOTE + Tomek Links | 0,733 | 0,250 | 0,220 | 0,290 | 0,460 | 0,940 |
| **Logistic Regression** | Pipeline Inicial | 0,495 | 0,059 | 0,030 | 0,430 | 0,684 | 0,560 |
| Undersampling | 0,627 | 0,108 | 0,060 | 0,570 | 0,850 | 0,690 |
| SMOTE | 0,505 | 0,063 | 0,030 | 0,430 | 0,706 | 0,590 |
| SMOTE + Tomek Links | 0,502 | 0,064 | 0,030 | 0,430 | 0,705 | 0,590 |
| **SVM** | Pipeline Inicial | 0,331 | 0,073 | 0,040 | 0,860 | 0,041 | 0,300 |
| Undersampling | 0,347 | 0,104 | 0,060 | 0,570 | 0,243 | 0,680 |
| SMOTE | 0,607 | 0,140 | 0,080 | 0,570 | 0,959 | 0,770 |
| SMOTE + Tomek Links | 0,604 | 0,138 | 0,080 | 0,570 | 0,977 | 0,770 |
| **XGBoost** | Pipeline Inicial | 0,775 | 0,400 | 0,380 | 0,430 | 0,701 | 0,960 |
| Undersampling | 0,780 | 0,096 | 0,050 | 0,860 | 0,992 | 0,100 |
| SMOTE | 0,734 | 0,174 | 0,120 | 0,290 | 0,999 | 0,910 |
| SMOTE + Tomek Links | 0,712 | 0,222 | 0,150 | 0,430 | 0,999 | 0,900 |

Fonte: Própria autoria.

Relembrando rapidamente as métricas utilizadas:

* ROC-AUC (área sob a curva ROC): mede a capacidade do modelo de distinguir entre classes (mineral vs. não mineral), sendo útil para avaliar o desempenho geral, porém não maximiza a taxa de TP;
* F1-Score: média harmônica entre precision e recall, sendo útil para avaliar desequilíbrio entre classes, como o presente caso, porém não é ótimo para priorizar maximização de TP;
* Precision (precisão): razão entre TP e total de predições positivas;
* Recall (revocação): razão entre TP e total de reais positivos;
* Threshold: limiar ótimo de probabilidades do modelo, ou seja, acima deste limiar de probabilidade, o ponto é classificado como ocorrência;
* Acurácia: razão entre predições corretas e total de amostras.

Considerando o problema de negócio (predição de minerais), há margem para a métrica de maximização. Por um lado, busca-se uma alta taxa de verdadeiros positivos (*true positive* – TP), ou seja, um alto recall. Todavia, considerando custos de um projeto de mineração, falsos positivos seriam altamente custosos, sendo a métrica precision mais adequada. Neste trabalho, optou-se pela segunda abordagem, na qual busca-se minimizar os falsos positivos, ou seja, alta precision.

Sendo assim, o modelo XGBoost com o pipeline inicial apresentou maior precision (0,38), sendo o modelo escolhido para continuação da análise.

A seguir, é apresentado o relatório de classificação elaborado pelo scikit learn com o método classification\_report, para o modelo XGBoost, sem e com a aplicação do limiar de probabilidades. Observa-se que a precisão se eleva de 0,38 para 0,50 com a utilização de tal limiar, trazendo ganhos a o modelo.

Figura 8 – Relatórios de classificação para ocorrências de ouro com e sem aplicação de limiar de probabilidade.

|  |
| --- |
| (a)  Tela de celular com texto preto sobre fundo branco  O conteúdo gerado por IA pode estar incorreto. |
| (b)  Calendário  O conteúdo gerado por IA pode estar incorreto. |

Fonte: Própria autoria.

A seguir, é apresentada a figura de ROC-AUC, a qual apresenta uma comparação entre as discriminações realizadas pelo modelo em azul, com um modelo aleatório (linha diagonal vermelha). No caso, o modelo em questão possui uma capacidade discriminatório superior a um modelo aleatório, com um AUC de 0,78, o qual é moderadamente bom, nos dizendo que o modelo tem 78% de chance de classificar corretamente um mineral corretamente.

Figura 9 – ROC-AUC para as ocorrências de ouro.

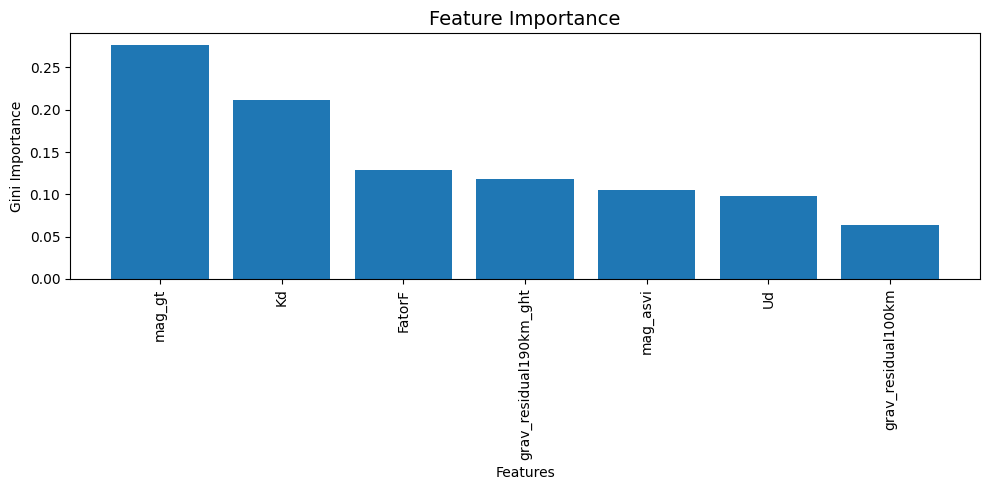
Gráfico, Gráfico de linhas

O conteúdo gerado por IA pode estar incorreto.

Fonte: Própria autoria.

A seguir é apresentada a importância de características (*feature importance*) do modelo. No caso, os modelos de árvore medem a contribuição média de cada feature para reduzir a impureza de Gini (i.e. redução do erro) em todas as árvores do modelo ou, de maneira simplificada, o que o modelo utilizou para as probabilidades/ocorrências. Observa-se que a variável mag\_gt se mostrou como a mais importante no cálculo, seguida pela variável Kd. Cabe destacar que a análise da metodologia de árvores é geral (toda amostra), e não considera exatamente a direção da influência (i.e. positiva ou negativa).

Figura 10 – Feature importance do modelo XGBoost para ocorrências de ouro.



Fonte: Própria autoria.

Buscou-se analisar também a importância dos features via biblioteca shap, a qual se baseia na Teoria dos Jogos de Shapley, atribuindo um valor de contribuição para cada feature por amostra, considerando todas as combinações possíveis de features. Aqui, o diferencial é que a análise é global (média do impacto absoluto) e local (amostra). A figura a seguir apresenta um gráfico de (a) barras e (b) violino, sendo que este apresenta tanto a importância do feature quanto sua direção. Como a análise é local e global, observa-se que determinada parte da amostra da variável mag\_gt, a mais importante também nesta metodologia, possui influência negativa, enquanto outra parte possui influência positiva.

Figura 11 – Feature importance com SHAP para ocorrências de ouro em gráfico de (a) barras e (b) violino.

|  |
| --- |
| (a)  Gráfico, Gráfico de barras  O conteúdo gerado por IA pode estar incorreto. |
| (b)  Gráfico  O conteúdo gerado por IA pode estar incorreto. |

Fonte: Própria autoria.

Por fim, após calculadas as probabilidades aplicando-se o threshold, foi possível criar uma figura iterativa com a biblioteca folium, a qual apresenta locais com minas em vermelho, ocorrências de ouro em laranja, e possíveis ocorrências de ouro com alta probabilidade advindas do modelo em amarelo. Observa-se que diversos pontos de ocorrência estão praticamente sobrepostos, resultante da grande proximidade entre pontos geográficos. Todavia, existe um ponto com alta probabilidade em destaque que foge às localizações com ouro já existente.

Figura 12 – Figura com localizações de minas (vermelho), ocorrências de ouro (laranja) e possíveis ocorrências de ouro com alta probabilidade (amarelo).

Mapa

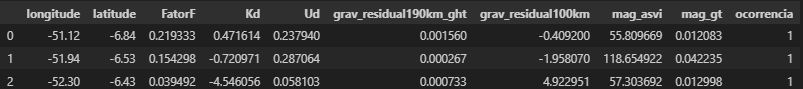
O conteúdo gerado por IA pode estar incorreto.

Fonte: Própria autoria.

## Ocorrências de cobre

A figura a seguir apresenta uma captura de tela do dataframe utilizado para predição de ocorrências de ouro:

Figura 13 – Dados utilizados para previsão de cobre.



Fonte: Própria autoria.

As figuras a seguir apresentam os passos aplicados para redução do desbalanceamento, conforme citado e apresentado anteriormente.

Figura 14 – (a) Área inicial e (b) área recortada com reamostragem para redução de desbalanceamento para ocorrências de cobre.

|  |
| --- |
| (a) |
| (b) |

Fonte: Própria autoria.

De maneira a ilustrar a redução, a base de dados final para modelagem é representada pela figura a seguir. Observam-se um total de 1.175 pontos, sendo 42 deles de presença de ouro, ou seja, 3,57% do dataset, uma grande elevação em relação ao mencionado anteriormente, de 0,4%.

Figura 15 – Contagem dos dados por ocorrência, sendo 1 a ocorrência de cobre.

Gráfico, Gráfico de barras, Histograma

O conteúdo gerado por IA pode estar incorreto.

Fonte: Própria autoria.

A tabela a seguir apresenta os resultados para as ocorrências de cobre. Nota-se que o modelo Random Forest com a pipeline oversampling (SMOTE) apresentou o melhor resultado de precision, sendo o modelo escolhido para a continuação da análise.

Tabela 2 – Métricas dos modelos por pipeline para ocorrências de cobre.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **Abordagem** | **ROC-AUC** | **F1-Score** | **Precision** | **Recall** | **Threshold** | **Acurácia** |
| **Random Forest** | Pipeline Inicial | 0,716 | 0 | 0 | 0 | 0,15 | 0,96 |
| Undersampling | 0,716 | 0,134 | 0,07 | 0,69 | 0,81 | 0,67 |
| SMOTE | 0,677 | 0,308 | 0,31 | 0,31 | 0,66 | 0,95 |
| SMOTE+Tomek | 0,656 | 0,296 | 0,29 | 0,31 | 0,56 | 0,95 |
| **Logistic Regression** | Pipeline Inicial | 0,687 | 0,114 | 0,06 | 0,62 | 0,57 | 0,65 |
| Undersampling | 0,705 | 0,101 | 0,06 | 0,62 | 0,59 | 0,6 |
| SMOTE | 0,676 | 0,113 | 0,06 | 0,62 | 0,565 | 0,64 |
| SMOTE+Tomek | 0,616 | 0,107 | 0,06 | 0,54 | 0,478 | 0,67 |
| **SVM** | Pipeline Inicial | 0,416 | 0,082 | 0,04 | 0,77 | 0,035 | 0,37 |
| Undersampling | 0,718 | 0,12 | 0,07 | 0,62 | 0,511 | 0,67 |
| SMOTE | 0,844 | 0,256 | 0,15 | 0,77 | 0,879 | 0,84 |
| SMOTE+Tomek | 0,824 | 0,235 | 0,15 | 0,62 | 0,955 | 0,85 |
| **XGBoost** | Pipeline Inicial | 0,722 | 0,1 | 0,14 | 0,08 | 0,018 | 0,95 |
| Undersampling | 0,696 | 0,095 | 0,05 | 0,77 | 0,999 | 0,46 |
| SMOTE | 0,747 | 0,245 | 0,17 | 0,46 | 0,961 | 0,9 |
| SMOTE+Tomek | 0,754 | 0,286 | 0,23 | 0,38 | 0,894 | 0,93 |

Fonte: Própria autoria.

De maneira similar às ocorrências de ouro, utilizou-se o método classification\_report para comparar os ganhos do modelo sem e com a aplicação do limiar de probabilidades. Observa-se que a precisão se eleva de 0,17 para 0,25 com a utilização de tal limiar, trazendo ganhos a o modelo. Aqui é importante fazer um comentário sobre a diferença de valores apresentados na tabela para a seleção do modelo e as saídas do relatório de classificação, visto que, enquanto a tabela é calculada com o pipeline utilizando validação cruzada (SMOTE é aplicado separadamente para cada subconjunto), aqui não se aplicou a validação cruzada utilizada na etapa anterior, ou seja, a distribuição de amostras sintéticas difere, levando a limites de decisão e calibrações de probabilidade ligeiramente diferentes.

Figura 16 – Relatórios de classificação para ocorrências de cobre com e sem aplicação de limiar de probabilidade.

|  |
| --- |
| (a)  Calendário  O conteúdo gerado por IA pode estar incorreto. |
| (b)  Calendário  O conteúdo gerado por IA pode estar incorreto. |

Fonte: Própria autoria.

A seguir, é apresentada a figura de ROC-AUC, onde observa-se que o modelo em questão possui uma capacidade discriminatório superior a um modelo aleatório, com um AUC de 0,75, o qual é moderadamente bom, nos dizendo que o modelo tem 75% de chance de classificar corretamente um mineral corretamente.

Figura 17 – ROC-AUC para as ocorrências de cobre.

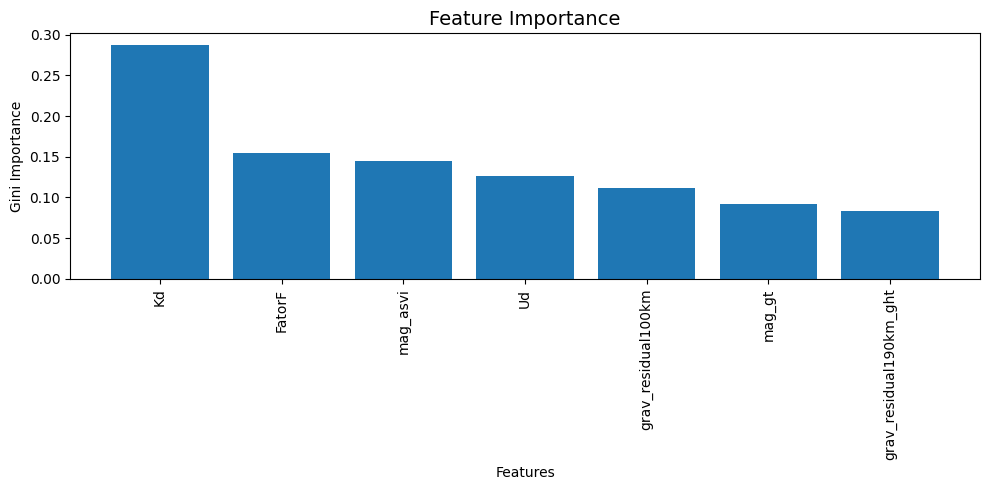
Gráfico, Gráfico de linhas

O conteúdo gerado por IA pode estar incorreto.

Fonte: Própria autoria.

A seguir é apresentado um gráfico com *feature importance* do modelo. Nota-se que a variável Kd se mostrou como a mais importante no cálculo, seguida pela variável FatorF, ambas do dataset de espectrometria. Outro ponto interessante é que a feature Kd apresenta uma importância muito superior às demais.

Figura 18 – Feature importance do modelo Random Forest para ocorrências de cobre.



Fonte: Própria autoria.

A análise SHAP resultou em uma variável de maior importância similar, Kd. De maneira similar à análise de ouro, a variável apresentou mais influência negativa no modelo do que positiva, todavia com um impacto inferior ao calculado pelo próprio modelo, como mencionado acima.

Figura 19 – Feature importance com SHAP para ocorrências de cobre em gráfico de (a) barras e (b) violino.

|  |
| --- |
| (a)  Gráfico, Gráfico de barras  O conteúdo gerado por IA pode estar incorreto. |
| (b)  Gráfico  O conteúdo gerado por IA pode estar incorreto. |

Fonte: Própria autoria.

Por fim, após calculadas as probabilidades aplicando-se o threshold, criou-se uma figura iterativa, a qual apresenta locais com minas em vermelho, ocorrências de cobre em marrom, e possíveis ocorrências de cobre com alta probabilidade advindas do modelo em coral. Observa-se que diversos pontos de ocorrência estão praticamente sobrepostos, resultante da grande proximidade entre pontos geográficos. Todavia, existes pontos com alta probabilidade em destaque que foge às localizações com cobre já existentes. Outro ponto é que, considerando a maior ocorrência natural de cobre em relação ao ouro, tal resultado é também visto na figura resultante a seguir.

Figura 20 – Figura com localizações de minas (vermelho), ocorrências de cobre (marrom) e possíveis ocorrências de cobre com alta probabilidade (coral).

Mapa

O conteúdo gerado por IA pode estar incorreto.

Fonte: Própria autoria.

# Referências

RUTECKI, Marcin. **Best techniques and metrics for Imbalanced Dataset.** Disponível em: https://www.kaggle.com/code/marcinrutecki/best-techniques-and-metrics-for-imbalanced-dataset. Acesso em: 18 maio 2025.