

É possível classificar
solos usando técnicas de
machine learning?

Aline Gallo De Mitri

Introdução ao Problema



CONTEXTO

Potencial de uso de um terreno



CLASSIFICAÇÃO DO SOLO

Avaliação em diferentes etapas



SUBJETIVIDADE

Processo individual e oneroso



TOMADA DE DECISÃO

Instrumento prévio que pode acelerar decisão

Recorte



MATO GROSSO

11 classes de solo do primeiro nível





Principal Objetivo

Criar um modelo que, a partir dos dados dos perfis de solo, permita prever a classe do solo de uma área na região do Mato Grosso

Dataset

- O dataset foi obtido no banco de dados do Sistema Brasileiro de Classificação de Solos (Embrapa)



Limpeza Inicial e Transformação de Variáveis

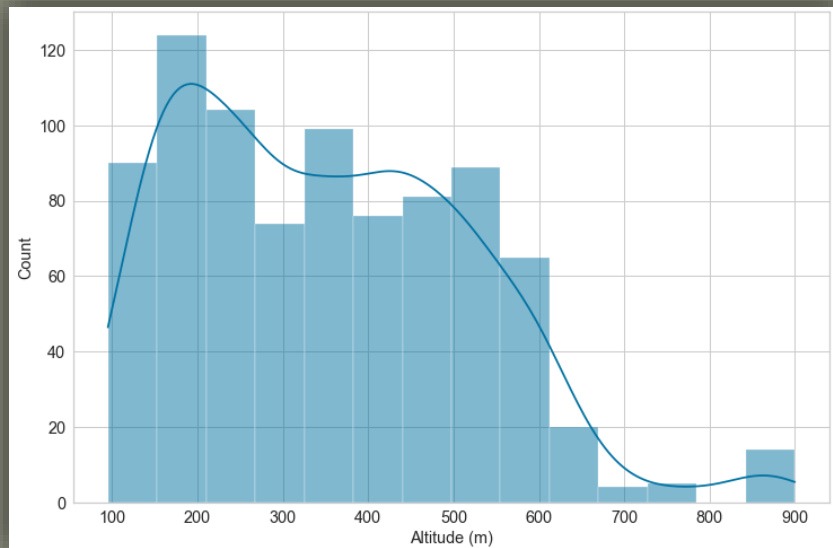
- 90 variáveis apresentavam todos os seus valores como indefinidos (“NaN”)
- Mudanças no tipo de algumas variáveis
- Alteração em valores com símbolo “<”

The background features a light gray base with large, organic, overlapping shapes in muted olive green and dusty rose. In the top left corner, there is a stylized illustration of a pine branch with needle-like leaves in a light gray tone. The title text is centered within the dusty rose shape.

Análise Exploratória

Variáveis Quantitativas

IDENTIFICAÇÃO



CARACTERÍSTICAS FÍSICAS

Composição granulométrica da terra fina

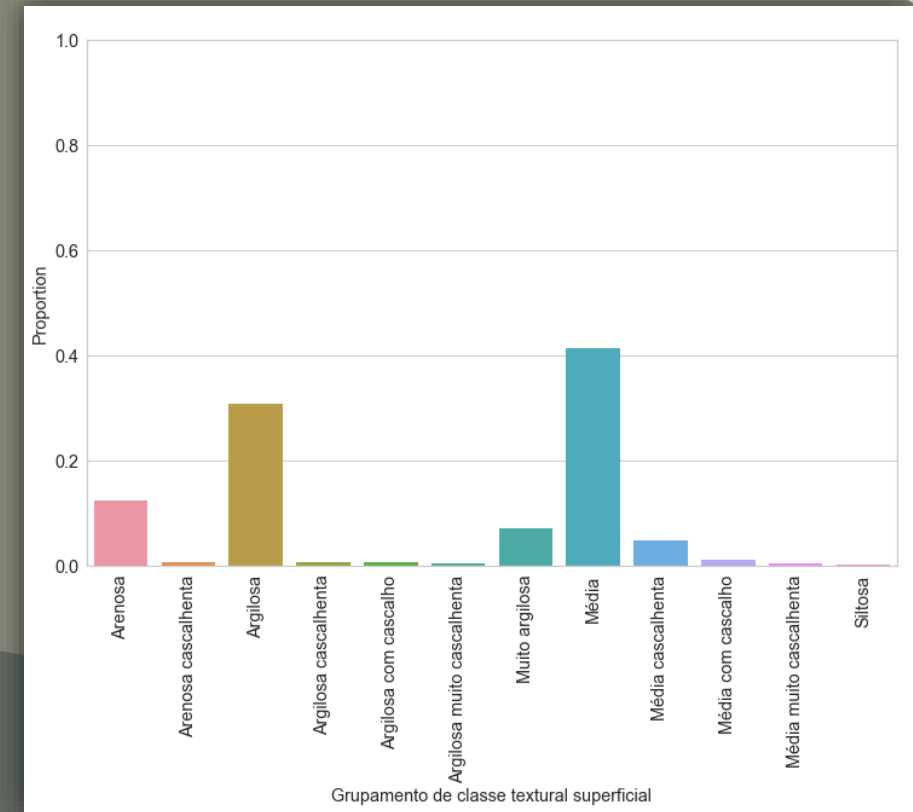
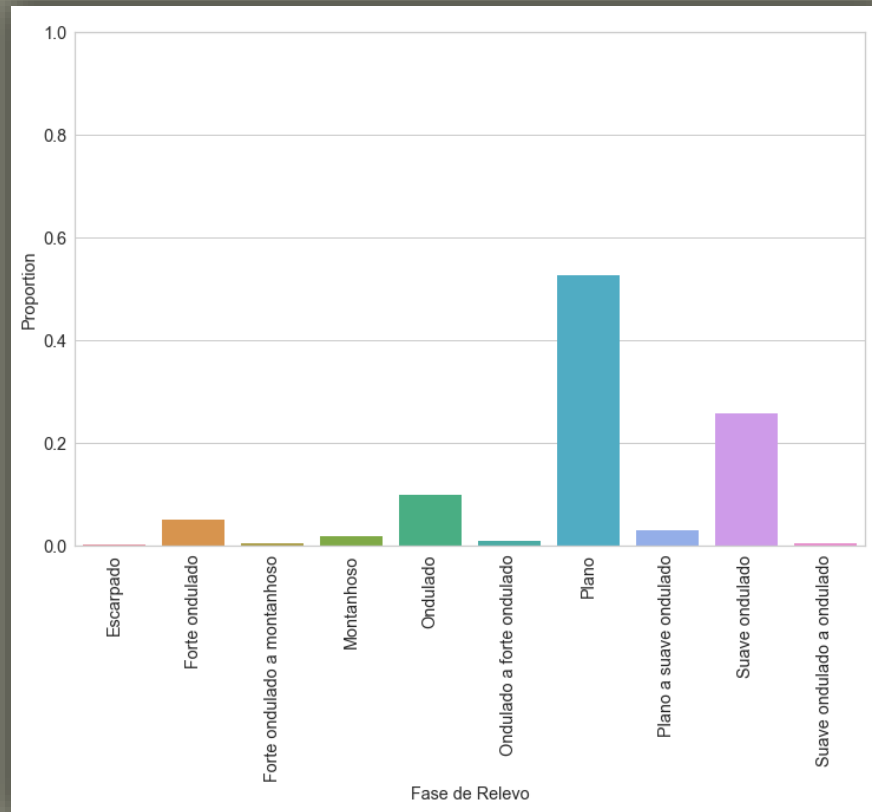
- Areia: 51%
- Argila: 31%
- Silte: 18%

Variáveis Quantitativas

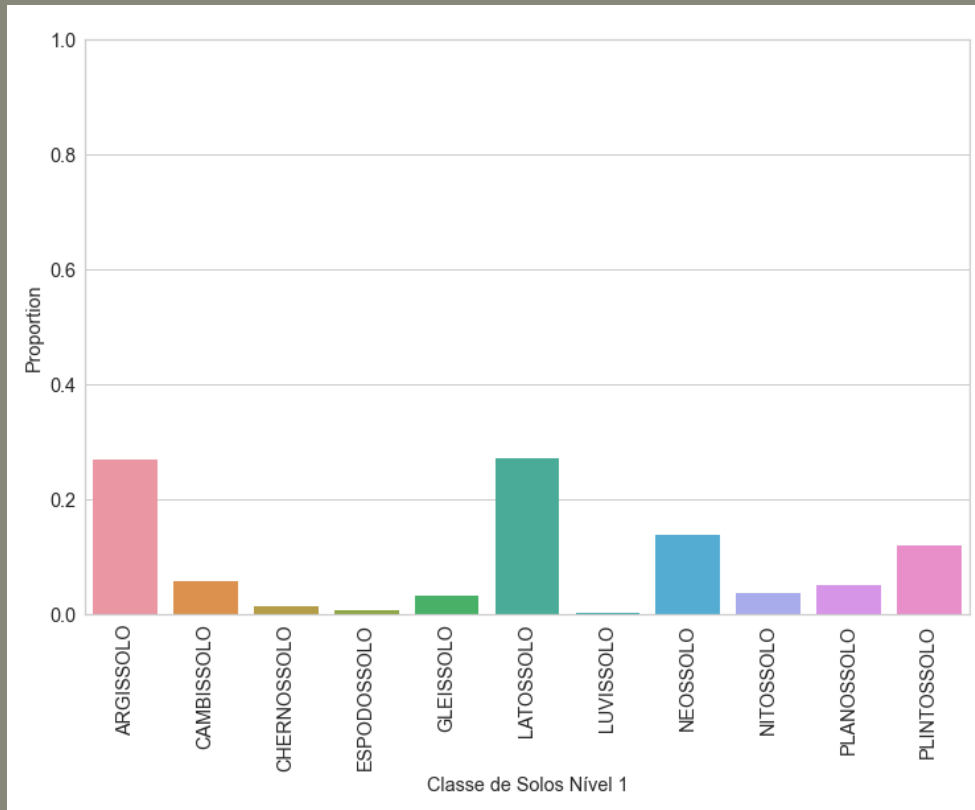
CARACTERÍSTICAS QUÍMICAS

Variável		Ca ²⁺	Mg ²⁺	Al ³⁺	Carbono	pH
Valor médio observado		2,07	0,72	0,68	8,11	5,38
Valor recomendado	Baixo	< 1,0	< 0,4	< 0,3	< 8,0	Entre 5,8 e 6,2
	Médio	1,0-2,0	0,4-0,8	0,3-0,8	8,0-14,0	
	Alto	> 2,0	> 0,8	> 0,8	>14,0	

Variáveis Qualitativas




Variável Resposta



Classe de Solos	Número de Perfis	Número da Classe nos Modelos
Argissolo	351	0
Cambissolo	75	1
Chernossolo	17	2
Espodossolo	9	3
Gleissolo	42	4
Latossolo	355	5
Luvissolo	4	6
Neossolo	179	7
Nitossolo	48	8
Planossolo	67	9
Plintossolo	155	10

Limpeza e Seleção de Variáveis

- Construção do modelo apenas com as variáveis quantitativas
- Remoção de outliers
- Transformação dos valores indefinidos para 0
- Transformação das classes da variável resposta em números
- Ajuste das escalas

The background features a light gray base with large, organic, overlapping shapes in muted olive green and dusty rose. A stylized fern frond is visible in the upper left corner. A thin white line curves across the bottom right.

Construção e Análise dos Modelos

Modelos de Classificação

Regressão Logística

Define uma relação linear entre as variáveis de entrada e a variável resposta, sendo esta variável modelada entre as diferentes classes. Este modelo foi utilizado em sua forma “One vs. Rest”, já que estamos lidando com um problema multiclasse.

Random Forest

Define diferentes conjuntos de regras a partir das características de uma amostra de dados para definir regras de classificação. Cada conjunto de regras é conhecido como árvore de decisão, por isso o nome “floresta aleatória”.

XGBoost

Modelo baseado na árvore de decisão, mas em que cada nó da árvore é um modelo de aprendizado “fraco”, o qual fornece informações para o minimizar os erros do modelo seguinte, até se obter um único modelo “forte”, como resultados mais robustos.

One vs. Rest


Decompõe o problema de classificação multiclasse em problemas binários em que uma classe é comparada com o todas as outras classes.

Support Vector Machine

Define fronteiras de decisão a partir das características dos dados. Cada fronteira separa uma classe da outra.

Avaliação dos Modelos

Modelo	Acurácia	Precisão	Sensibilidade
Regressão Logística	0,61	0,62	0,61
Random Forest	0,68	0,68	0,68
XGBoost	0,70	0,70	0,70
One vs. Rest	0,54	0,65	0,54
Support Vector Machine	0,66	0,65	0,66

The background features a light gray base with large, organic, overlapping shapes in muted olive green and dusty rose. A stylized fern frond is visible in the upper left corner. A thin white line curves across the bottom right.

Implicações e Próximos Passos

Resumo

Todos os modelos apresentaram um desempenho razoável de previsão da classe de solos dos perfis de teste.

O melhor modelo encontrado foi o **XGBoost**, com acurácia e precisão de 70%.



Próximos Passos

Voltar a análise de dados e realizar uma avaliação mais detalhada das variáveis quantitativas para melhorar o desempenho do modelo.

Estudo mais aprofundado para incluir variáveis qualitativas, principalmente perfil do horizonte.

Após a concretização do modelo, apresentar uma solução web em que os usuários poderiam consultar um mapa interativo do estado de Mato Grosso com as regiões de cada classe.





Obrigada!

Aline Gallo De Mitri

alinegmitri@gmail.com