IX Conferência Brasileira sobre Estabilidade de Encostas COBRAE 17 A 21 DE NOVEMBRO
GEOSUL XIV Simpósio de Prática de Engenharia Geotécnica da Região Sul
BENTO GONÇALVES - RS
2025

Uma análise abrangente de métodos de aprendizado de máquina para classificação de solos com base em parâmetros *in situ*

Helena Paula Nierwinski
Professora, Universidade Federal de Santa Catarina, Joinville, Brasil, helena.paula@ufsc.br

Ricardo José Pfitscher
Professor, Universidade Federal de Santa Catarina, Joinville, Brasil, ricardo.pfitscher@ufsc.br

Alberthus Koops Lordello
Bacharel em Ciência e Tecnologia, Universidade Federal de Santa Catarina, Joinville, Brasil, alberthus.lordello@gmail.com

RESUMO: A análise, caracterização e classificação dos solos são etapas fundamentais para diversas aplicações da engenharia geotécnica. Os ensaios laboratoriais tradicionais, embora precisos, costumam ser dispendiosos, demandar tempo e fornecer resultados limitados a pontos específicos de amostragem, dificultando a extrapolação para áreas mais amplas. O ensaio de piezocone (CPT) tem se consolidado como uma alternativa viável, devido ao seu menor custo, rápida execução e capacidade de fornecer dados consistentes em regiões geograficamente próximas. Este estudo utiliza técnicas de aprendizado de máquina (AM) para aprimorar a classificação dos solos, por meio do treinamento de algoritmos com um extenso conjunto de dados obtidos a partir de ensaios CPT. Diversos modelos de AM foram avaliados, sendo que o algoritmo de melhor desempenho atingiu uma acurácia de classificação de 92,9%. Os resultados evidenciam o potencial do aprendizado de máquina para otimizar os processos de classificação de solos, reduzindo significativamente os custos e ampliando a escalabilidade em investigações geotécnicas de larga escala.

PALAVRAS-CHAVE: Classificação de solos, CPTu, Análise de solos, Inteligência artificial, Investigação geotécnica.

ABSTRACT: The analysis, characterization, and classification of soils are fundamental steps for various geotechnical engineering applications. Traditional laboratory tests, although accurate, are often costly, time-consuming, and limited to discrete sampling points, which hinders the extrapolation of results to larger areas. The piezocone penetration test (CPTu) has emerged as a viable alternative due to its lower cost, rapid execution, and ability to provide consistent data across geographically close regions. This study employs machine learning (ML) techniques to enhance soil classification by training algorithms on an extensive dataset derived from CPTu tests. Several ML models were evaluated, with the best-performing algorithm achieving a classification accuracy of 92.9%. The results highlight the potential of machine learning to optimize soil classification processes, significantly reduce costs, and improve scalability in large-scale geotechnical investigations.

KEYWORDS: Soil classification, CPTu, Soil analysis, Artificial intelligence, Geotechnical investigation.

1 INTRODUÇÃO

A classificação dos solos é uma etapa essencial em projetos de engenharia geotécnica, sendo tradicionalmente realizada a partir da identificação de suas características físicas e propriedades intrínsecas (Souza Pinto, 2006). De acordo com Aydin (2023), essas propriedades são utilizadas para definir parâmetros que subsidiam a caracterização e a classificação dos solos, geralmente por meio de ensaios laboratoriais sobre amostras coletadas em campo. No entanto, parte desses parâmetros também pode ser estimada com base em resultados de ensaios *in situ*, permitindo uma avaliação mais abrangente. O agrupamento dos solos com base em

propriedades similares torna-se, assim, um processo chave para o dimensionamento de fundações e o desenvolvimento de projetos geotécnicos mais seguros e eficientes.

Apesar da precisão dos ensaios laboratoriais, eles são frequentemente dispendiosos, exigem equipamentos especializados e dependem de equipes técnicas qualificadas para análise dos resultados (Puri, 2018). Além disso, conforme observado por Nierwinski *et al*. (2023), há um desafio inerente na extrapolação dos dados obtidos a partir de amostras pontuais para toda a área de um projeto, o que limita sua representatividade espacial.

Como alternativa ou complemento aos ensaios laboratoriais, o ensaio de piezocone (CPTu) tem sido amplamente utilizado em investigações geotécnicas. Segundo Bhattacharya e Solomatine (2006), o CPTu consiste na cravação contínua de um cone metálico no solo, em velocidade constante, permitindo a obtenção de parâmetros como a resistência de ponta (q_t), o atrito lateral (f_s) e a poropressão (u_2). Este método fornece perfis contínuos de propriedades do solo com alta resolução espacial, eliminando a necessidade de envio de amostras ao laboratório e viabilizando interpretações mais rápidas e econômicas.

Além dos parâmetros obtidos diretamente do CPTu, variáveis complementares como a densidade real dos grãos (G) e o peso específico do solo (γ), determinadas em laboratório, são frequentemente utilizadas na caracterização geotécnica. Conforme Nierwinski *et al*. (2023) e Menegaz *et al*. (2022), a densidade real é uma propriedade intrínseca que auxilia na distinção entre diferentes tipos de solo, enquanto o peso específico é fundamental para cálculos de tensões e dimensionamento de estruturas de fundação.

Nas últimas décadas, técnicas de aprendizado de máquina (AM) têm sido aplicadas com sucesso à classificação de solos, conforme destacado por Chandan (2018). Essas abordagens incluem desde a previsão de áreas com risco de instabilidade até classificações específicas para fins agrícolas e de engenharia, com crescente acurácia e capacidade de generalização.

Neste estudo, investiga-se a aplicação de diferentes algoritmos de aprendizado de máquina na classificação de solos, utilizando como variáveis preditoras os parâmetros obtidos em ensaios CPTu (q_t, f_s, u_2) e os parâmetros laboratoriais G e γ. Um diferencial importante deste trabalho é o uso de um conjunto de dados robusto, composto por amostras de diferentes regiões geográficas e tipos de solo, o que permite avaliar o desempenho dos modelos em um cenário mais generalizado e representativo.

2 CONCEITOS TEÓRICOS

2.1 Transformação Yeo-Johnson

A transformação de Yeo-Johnson é uma generalização da transformação Box-Cox, adequada para variáveis com valores nulos ou negativos, comumente encontradas em dados geotécnicos com distribuição log-normal. Essa técnica permite estabilizar a variância e aproximar os dados de uma distribuição gaussiana, sendo particularmente útil para melhorar o desempenho de algoritmos de aprendizado de máquina em conjuntos de dados com presença de outliers e escalas heterogêneas (Riani *et al*., 2023).

2.2 Algoritmos de Classificação Utilizados

No presente estudo, foram avaliados diferentes algoritmos de aprendizado de máquina com o objetivo de classificar tipos de solo a partir de parâmetros in situ e laboratoriais. Abaixo, apresenta-se uma síntese dos modelos selecionados:

- *K-Nearest Neighbors* (*KNN*): Algoritmo baseado em proximidade, que atribui a uma nova amostra a classe predominante entre seus k vizinhos mais próximos. Sua aplicação é sensível à escala dos atributos e ao balanceamento do conjunto de dados (Müller & Guido, 2016);
- Regressão Logística Multinomial: Modelo probabilístico que utiliza a função logística para estimar a probabilidade de uma amostra pertencer a uma determinada classe. Apesar de linear, apresenta bom desempenho quando os dados são normalizados (Albon, 2018);
- Gaussian *Naive Bayes:* Método probabilístico que assume independência entre os atributos e distribuição gaussiana das variáveis. É eficiente para problemas com classes bem separadas e atributos com baixa correlação (Albon, 2018).
- Árvore de Decisão: Modelo não linear que realiza partições sucessivas do espaço de atributos com base

na minimização da impureza Gini, gerando regras de decisão interpretáveis. É eficaz para conjuntos de dados heterogêneos e com interações não lineares entre variáveis (Bonaccorso, 2017).

- Floresta Aleatória (*Random Forest*): Conjunto de árvores de decisão treinadas com subconjuntos aleatórios de dados e atributos (*bagging*), visando reduzir o *overfitting* e melhorar a robustez do modelo. É amplamente utilizada em problemas de classificação em geotecnia devido à sua alta acurácia e estabilidade (Müller & Guido, 2016).

2.3 Validação Cruzada

A validação cruzada (*k-fold cross-validation*) é uma técnica amplamente adotada para avaliação da performance de modelos de aprendizado de máquina. O conjunto de dados é particionado em *k* subconjuntos, sendo cada um utilizado como teste em uma rodada, enquanto os demais servem como treino. Essa abordagem permite obter estimativas mais confiáveis de desempenho, especialmente em bases de dados geotécnicos com classes desbalanceadas ou amostras limitadas (Burman, 1989).

3 METODOLOGIA

3.1 Base de dados

O conjunto de dados utilizado neste estudo foi o mesmo utilizado no estudo de Nierwinski *et al*. (2023), o qual contém 1.862 amostras, representativas de dez diferentes classes de solo, coletadas em diversas regiões geográficas e com ampla variação de características geotécnicas. As variáveis incluídas no *dataset* compreendem parâmetros obtidos em ensaios de piezocone (resistência de ponta - q_t, atrito lateral - f_s, poropressão - u_2 e razão de atrito - *Rf*) e ensaios laboratoriais (densidade real dos grãos - *G* e peso específico - γ). A distribuição dos tipos de solo que compõe a base de dados pode ser observada na Figura 1.

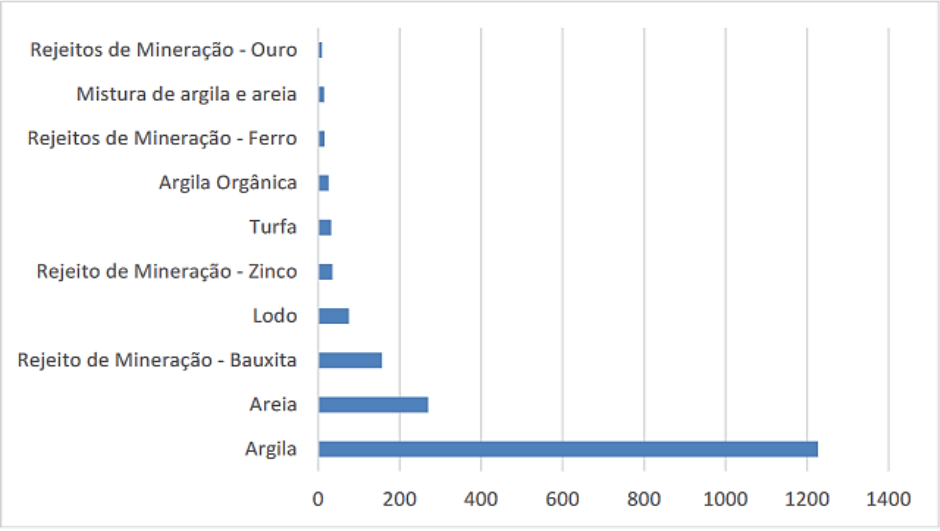

Figura 1. Distribuição dos tipos de solo no *dataset*.

3.2 Análise Exploratória de Dados

Antes da aplicação dos modelos de aprendizado de máquina, foi realizada uma análise exploratória com o objetivo de compreender a distribuição dos dados, identificar *outliers* e avaliar a relação entre os atributos e os tipos de solo. Foram utilizados histogramas, gráficos de dispersão e estimativas de densidade *kernel* (KDE) para cada variável contínua, destacando possíveis assimetrias e padrões log-normais. A análise gráfica também auxiliou na definição das etapas de pré-processamento.

3.3 Pré-processamento dos Dados

A etapa de pré-processamento foi estruturada com o objetivo de otimizar o desempenho dos algoritmos de classificação. As seguintes etapas foram aplicadas:

- Tratamento de *Outliers*: Devido à sensibilidade de modelos lineares aos valores extremos, foi adotada uma abordagem conservadora, excluindo apenas os 2% de valores mais extremos para cada atributo (com exceção das variáveis G e γ, que mostraram ser importantes para a diferenciação entre classes). A decisão foi baseada na análise exploratória, com o intuito de preservar informações relevantes em classes minoritárias;
- Normalização com Yeo-Johnson: Como os atributos apresentaram distribuição semelhante à log-normal, utilizou-se a transformação de Yeo-Johnson para normalização dos dados. Esse método melhora a distribuição estatística e a convergência dos modelos, especialmente os lineares (Müller & Guido, 2016);
- Balanceamento com SMOTE: Para lidar com o desbalanceamento entre as classes de solo, foi testada a técnica de *Synthetic Minority Over-sampling Technique* (SMOTE), que gera novas amostras sintéticas das classes minoritárias com base na interpolação entre vizinhos próximos. Essa abordagem visa enriquecer o conjunto de dados de treinamento com variações realistas, favorecendo o aprendizado de regras classificatórias mais robustas (Bruce *et al*., 2020).

3.4 Modelagem e Algoritmos

Para a tarefa de classificação dos solos, foram avaliados cinco algoritmos de aprendizado de máquina, selecionados por sua diversidade de abordagem e desempenho consolidado em problemas de classificação supervisionada: Árvore de Decisão (*Decision Tree*), Floresta Aleatória (*Random Forest*), *K*-Vizinhos Mais Próximos (*K-Nearest Neighbors* - KNN), Regressão Logística Multinomial e *Naive Bayes* com distribuição Gaussiana.

Cada modelo foi treinado e testado em diferentes configurações, com e sem transformação de escala, e com ou sem aplicação do SMOTE. A avaliação da performance considerou métricas como acurácia, precisão, revocação (*recall*) e *f1-score*, utilizando validação cruzada *k-fold* (k=5) para garantir robustez estatística dos resultados.

4 RESULTADOS E DISCUSSÕES

4.1 Análise Exploratória dos Dados

A análise exploratória revelou importantes características estatísticas dos dados utilizados. Conforme ilustrado na Figura 2, os atributos obtidos a partir do ensaio CPTu (q_t, f_s, u e Rf) apresentaram distribuição assimétrica, com tendência log-normal e caudas longas — comportamento típico de parâmetros geotécnicos naturais, refletindo a heterogeneidade dos solos investigados. Este tipo de distribuição justifica o uso de técnicas de transformação como Yeo-Johnson para adequação dos dados aos modelos estatísticos.

Outro aspecto relevante foi observado na análise da dispersão entre os parâmetros obtidos em laboratório, G (densidade real dos grãos) e γ (peso específico natural do solo). Os pontos extremos dessas variáveis mostraram-se importantes para a distinção entre diferentes classes de solo, sendo indicativos de solos orgânicos, materiais de baixa compacidade ou rejeitos com comportamento atípico. Portanto, a exclusão destes valores como *outliers* poderia comprometer a representação da variabilidade natural do subsolo e mascarar transições litológicas relevantes para a engenharia geotécnica.

4.2 Estratégias de Pré-processamento

Com base na análise estatística inicial, adotou-se uma abordagem conservadora para o tratamento de *outliers*, restringindo a exclusão a apenas 2% das amostras mais extremas nas variáveis derivadas do CPTu.

As variáveis G e γ foram mantidas integralmente, assegurando a preservação de características intrínsecas de solos com comportamento geotécnico mais complexo.

A aplicação da transformação de Yeo-Johnson mostrou-se adequada, principalmente para melhorar a performance dos modelos lineares. Já o balanceamento de classes com o algoritmo SMOTE foi testado, mas os resultados finais indicaram ganho marginal ou inexistente na maioria dos casos. As tabelas 1 e 2 apresentam os valores estatísticos antes e depois do tratamento, confirmando a estabilidade do conjunto de dados após o pré-processamento, sem perda significativa de representatividade.

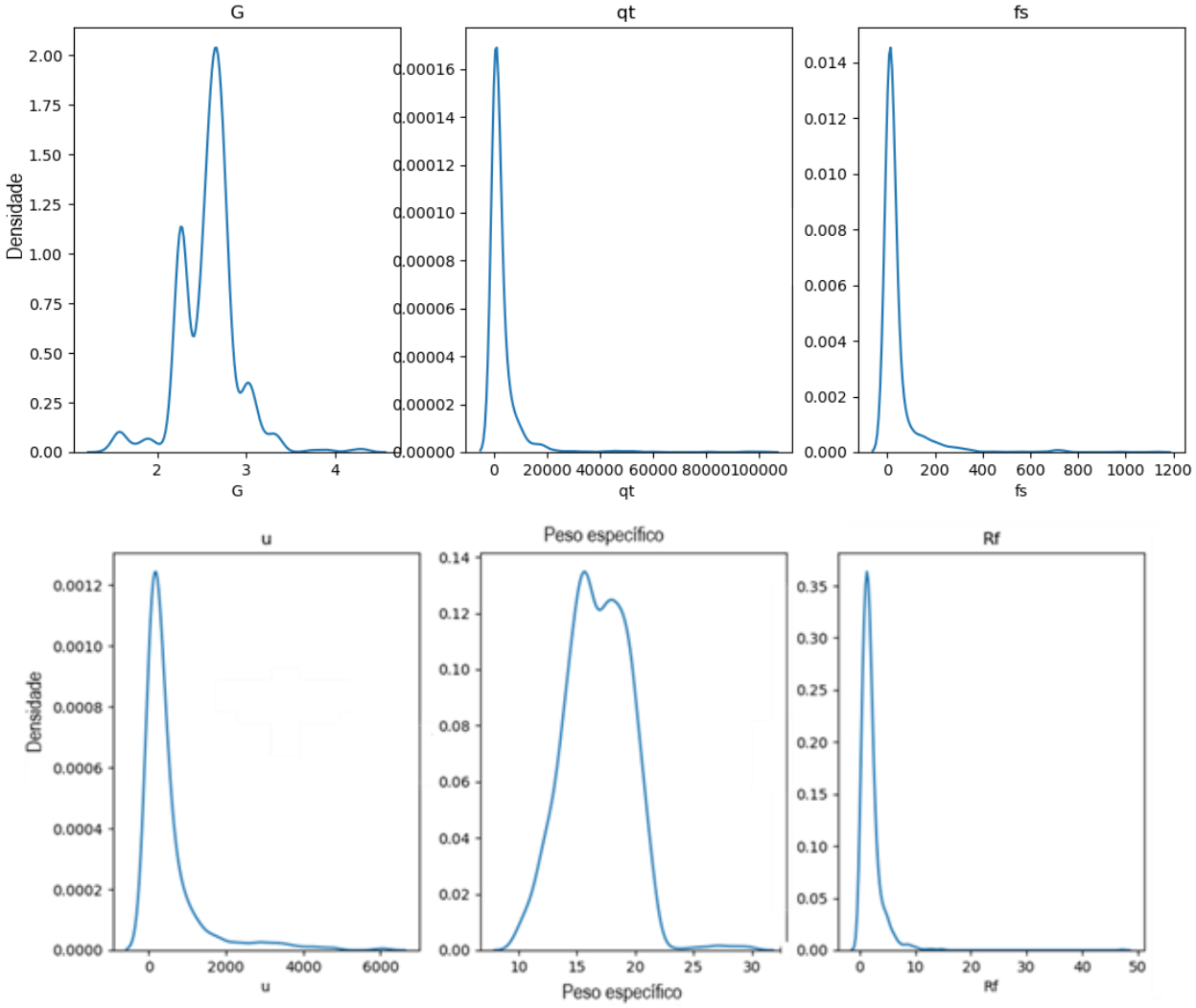

Figura 2. Distribuição KDE por atributo

Tabela 1. Distribuição estatística dos parâmetros antes do tratamento de *outlier*

	G	q_t (kPa)	f_s (kPa)	u_2 (kPa)	γ (kN/m³)	Rf
Amostras	1.488	1.488	1.488	1.488	1.488	1.488
Média	2,58	3.039,97	39,80	525,56	16,76	2,00
Desv. Pad.	0,33	7.393,01	87,66	812,53	2,75	2,06
min	1,45	30,37	0,00	0,00	9,84	0,00
25%	2,29	394,98	6,54	102,62	14,90	0,95
50%	2,62	862,62	13,11	251,58	16,70	1,58
75%	2,70	2.650,71	31,23	535,25	18,77	2,41
max	4,34	102.000,00	1.125,10	6.078,00	29,89	47,14

Tabela 2. Distribuição estatística dos parâmetros depois do tratamento de *outlier*

	G	q_t (kPa)	f_s (kPa)	u_2 (kPa)	γ (kN/m³)	Rf
Amostras	1.338	1.338	1.338	1.338	1.338	1.338
Média	2,58	2.358,61	32,56	511,16	16,77	1,90
Desv. Pad.	0,32	3.783,81	52,23	699,67	2,665	1,375
min	1,45	106,61	1,49	0,00	9,84	0,21
25%	2,28	427,27	7,12	119,45	15,02	0,98
50%	2,62	883,56	13,38	266,48	16,76	1,61
75%	2,70	2.505,99	30,18	555,88	18,68	2,36
max	4,34	34.507,38	373,49	4.155,00	29,89	9,06

4.3 Avaliação dos Modelos de Classificação

Após o pré-processamento, os modelos de aprendizado de máquina foram treinados e testados sob diferentes configurações. A Tabela 3 apresenta os resultados da acurácia para os modelos com e sem normalização por Yeo-Johnson. Verificou-se melhora significativa em modelos sensíveis à escala, como o KNN, regressão logística e Gaussian Naive Bayes, o que confirma a necessidade de normalização em bancos de dados geotécnicos com variabilidade elevada.

A Tabela 3 resume a performance dos algoritmos com base em validação cruzada (k=5), considerando as métricas de acurácia, revocação (*recall*) e *f1-score*. A Floresta Aleatória apresentou o melhor desempenho geral, com *f1-score* elevado e robustez frente à variabilidade dos dados. Esse modelo se destaca por sua capacidade de capturar não linearidades e interações entre variáveis, características comuns em sistemas geotécnicos com múltiplos fatores intervenientes.

Em contraste, os modelos lineares apresentaram desempenho inferior, mesmo após ajuste de escala, o que reforça a inadequação de suposições lineares em ambientes geotécnicos com solos heterogêneos, camadas intermediárias ou materiais de transição.

Tabela 3. Comparação da mudança de escala Yeo-Johnson

	Escalado	**Inalterado**	**Diferença**
Árvore de decisão	90,59	90,86	-0,27
Floresta aleatória	93,28	93,01	0,27
KNN	89,25	72,85	16,40
Regressão logística	88,17	74,19	13,98
Naive Baies Gausiana	82,53	59,95	25,58

Tabela 4. Validação cruzada da performance de cada modelo

	Acurácia (%)	***Recall* (%)**	***F1-score* (%)**
Árvore de decisão	89,3	86,9	86,6
Floresta aleatória	92,9	90,8	87,3
KNN	83,0	71,6	75,6
Regressão logística	67,9	62,5	66,4
Naive Baies Gausiana	70,3	78,4	79,1

4.4 Interpretação da Matriz de Confusão

A Figura 3 apresenta a matriz de confusão gerada pelo modelo de Floresta Aleatória, considerado o mais eficaz neste estudo. Os maiores índices de erro ocorreram entre as classes argila, areia, lodo e argila orgânica, cuja distinção, inclusive em análises convencionais, é desafiadora devido à sobreposição de propriedades físicas e proximidade de valores em parâmetros como q_t e f_s.

Apesar disso, os solos mais bem classificados foram aqueles com características mais marcantes, como os rejeitos de mineração (bauxita) e as argilas típicas, que apresentaram precisão superior a 90%. A acurácia

geral do modelo para os dados validados foi de 92,9%, valor expressivo considerando-se a diversidade de origem das amostras e a complexidade do banco de dados.

Esses resultados reforçam o potencial da inteligência artificial como ferramenta auxiliar em investigações geotécnicas, principalmente para ampliar a interpretação de dados in situ em projetos com restrições de tempo, orçamento ou com necessidade de análises em grande escala.

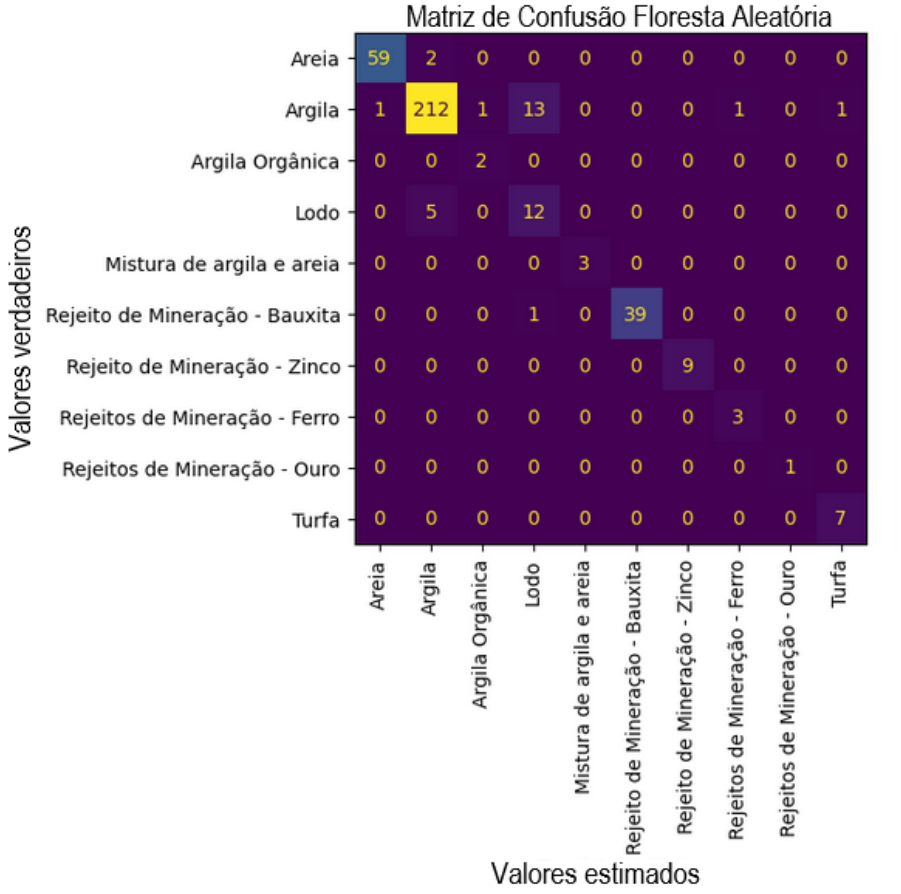

Figura 3. Matriz de confusão floresta aleatória

5 CONCLUSÕES

A classificação de solos é uma etapa essencial nos processos de investigação geotécnica, sendo determinante para o dimensionamento seguro e eficiente de fundações e estruturas de contenção. Este estudo demonstrou que a aplicação de algoritmos de aprendizado de máquina, alimentados com dados obtidos a partir de ensaios CPTu e parâmetros laboratoriais (G e γ), constitui uma abordagem promissora para a automatização e otimização da classificação de solos em larga escala. Uma possível limitação a ser apontada é que os modelos de classificação dependem dos valores G e γ, obtidos em laboratório. Contudo, como demonstrado em Nierwinski *et al*. (2023), é possível estimar os valores de peso específico utilizando os parâmetros de campo e os valores de G seguem um comportamento previsível (referência), tornando os modelos de classificação viáveis para aplicação em campo.

As principais contribuições desta pesquisa podem ser sintetizadas da seguinte forma:

i. Os algoritmos baseados em árvores, especialmente a Floresta Aleatória, apresentaram desempenho superior aos modelos lineares testados, com maior robustez frente à variabilidade dos dados e à complexidade intrínseca dos solos;
ii. A aplicação da transformação Yeo-Johnson mostrou-se benéfica apenas para os modelos lineares, melhorando sua capacidade de generalização. Para modelos não lineares, seu efeito foi marginal;

iii. O modelo desenvolvido não apresentou restrições quanto à localização geográfica, uma vez que foi treinado com amostras de diferentes regiões e tipos de solo, demonstrando bom potencial de generalização para aplicações geotécnicas diversas;
iv. A análise da matriz de confusão indicou que os principais erros de classificação ocorreram entre solos de características físicas semelhantes, como turfa, areia e argila. Já os demais tipos de solo, como os rejeitos de mineração, foram classificados com alta acurácia.

Esses resultados destacam o potencial de integração entre métodos de inteligência artificial e dados *in situ* na prática da engenharia geotécnica, promovendo maior escalabilidade e eficiência nos processos de investigação do subsolo.

REFERÊNCIAS BIBLIOGRÁFICAS

Albon, C. (2018). *Machine learning with python cookbook: Practical solutions from preprocessing to deep learning*. O'Reilly Media, Inc.

Aydın, Y., Işıkdağ, Ü., Bekdaş, G., Nigdeli, S. M., & Geem, Z. W. (2023). Use of Machine Learning Techniques in Soil Classification. *Sustainability*, 15(3), 2374.

Bhattacharya, B., Solomatine, D. P. (2006). Machine learning in soil classification. *Neural networks*, v. 19, n. 2, p. 186-195.

Bonaccorso, G. (2017). *Machine learning algorithms*. Packt Publishing Ltd. 155-158 e 167 p.

Burman, P. (1989). A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and the Repeated Learning-Testing Methods. *Biometrika*, 76(3), 503–514.

Bruce, P., Bruce, A.,Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.

Chandan, T. R. (2018). Recent trends of machine learning in soil classification: A review. *Int. J. Comput. Eng. Res*, v. 8, p. 25-33.

Menegaz T., Odebrecht E., Nierwinski H. P., Schnaid F. (2022). Soil unit weight prediction from CPTs for soils and mining tailings, in: *Cone Penetration Testing 2022*, CRC Press, 566–569.

Müller, A. C., Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.".

Nierwinski, H. P., Pfitscher, R. J., Barra, B. S., Menegaz, T., Odebrecht, E. (2023). A practical approach for soil unit weight estimation using artificial neural networks, *Journal of South American Earth Sciences*, V.131, 104648.

Puri, N., Prasad, H. D.,Jain, A. (2018). Prediction of geotechnical parameters using machine learning techniques. *Procedia Computer Science*, v. 125, p. 509-517.

Riani, M., Atkinson, A. C., Corbellini, A. (2023). Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression. *Statistical Methods & Applications*, v. 32, n. 1, p. 75-102.

Sousa Pinto, C. (2006). *Curso básico de mecânica dos solos*. 3ª ed. São Paulo: Oficina de Textos.