XXII Congresso Brasileiro de Mecânica dos Solos e Engenharia Geotécnica
XI Simpósio Brasileiro de Engenheiros Geotécnicos Jovens
24 a 28 de agosto de 2026 – Brasília/DF

DOI: 10.47094/COBRAMSEG2026.2/22

Otimização da Interpretação de Ensaios CPTU com Modelo de Aprendizado de Máquina: Uma Prova de Conceito

Pedro Gomes dos Santos Pereira
Engenheiro Geotécnico, JETbr, Rio de Janeiro, Brasil, ppereira@jetbr.eng.br

Fabio Marinho da Silva
Gerente de Engenharia, JETbr, Belo Horizonte, Brasil, fsilva@jetbr.eng.br

Nuno Alexandre Braz da Silva
Diretor de Engenharia, JETbr, Lisboa, Portugal, nsilva@jetsj.com

RESUMO: A interpretação de perfis geotécnicos a partir de ensaios de piezocone (CPTu) é essencial para o refinamento de perfis geológicos obtidos por sondagens, porém apresenta limitações decorrentes da subjetividade do analista e da dificuldade de ponderar múltiplas variáveis simultaneamente. Essa condição pode gerar inconsistências na definição dos horizontes geomecânicos, sobretudo em depósitos heterogêneos, como solos residuais. Este trabalho propõe a aplicação de técnicas de machine learning para automatizar e aprimorar a segmentação e parametrização de horizontes geomecânicos com base em dados de CPTu, reduzindo a subjetividade e integrando múltiplos parâmetros no processo decisório. Utilizaram-se dados de 32 ensaios, com classificação inicial baseada em condicionais pré-estabelecidas, contemplando Solo Residual Tipo A (Mole), Saprólito Tipo A, Solo Residual Tipo B (Fofo) e Saprólito Tipo B. A metodologia emprega um modelo supervisionado Random Forest, treinado com seis variáveis: poropressão hidrostática (u_0), poropressão (u_2), resistência de ponta (q_t), índice de comportamento (I_b), razão de atrito (F_r) e parâmetro de poropressão (B_q). Aplicou-se ainda suavização por janela móvel para reduzir hipersegmentação. Os resultados sugerem potencial ganho em produtividade, repetibilidade e eficiência, além de evidenciarem diferenças em relação à classificação inicial, indicando que a abordagem pode contribuir como ferramenta de suporte à decisão geotécnica.

PALAVRAS-CHAVE: CPTu, Machine Learning, Random Forest, Solo Residual

ABSTRACT: The interpretation of geotechnical profiles from piezocone tests (CPTu) is essential for refining geological profiles obtained from borehole investigations; however, it presents limitations arising from analyst subjectivity and the difficulty of simultaneously weighting multiple variables. This condition may lead to inconsistencies in the definition of geomechanical horizons, especially in heterogeneous deposits such as residual soils. This study proposes the application of machine learning techniques to automate and enhance the segmentation and parameterization of geomechanical horizons based on CPTu data, reducing subjectivity and integrating multiple parameters into the decision-making process. Data from 32 tests were used, with initial classification based on predefined conditional criteria, including Residual Soil Type A (Soft), Saprolite Type A, Residual Soil Type B (Loose), and Saprolite Type B. The methodology employs a supervised Random Forest model trained with six variables: hydrostatic pore pressure (u_0), pore pressure (u_2), cone tip resistance (q_t), behavior index (I_b), friction ratio (F_r), and pore pressure parameter (B_q). A moving window smoothing technique was also applied to reduce hyper-segmentation. The results suggest potential gains in productivity, repeatability, and efficiency, and highlight differences compared to the initial classification, indicating that the approach may contribute as a decision-support tool in geotechnical practice.

KEYWORDS: CPTu, Machine Learning, Random Forest, Residual Soil

1 INTRODUÇÃO

Na engenharia de barragens, a investigação geotécnica constitui etapa fundamental para a caracterização das fundações e materiais de construção, condicionante crítico à segurança e ao desempenho da estrutura. O ensaio de piezocone (CPTu), por se tratar de um ensaio confiável de alta repetibilidade e padronização, consolidou-se como ferramenta de alto refinamento para investigação in situ, oferecendo medições praticamente contínuas de resistência de ponta, atrito lateral e poropressão, as quais permitem inferir não apenas a estratigrafia, mas também parâmetros de resistência, deformabilidade e permeabilidade dos solos.

Apesar da interpretação de perfis geotécnicos a partir de dados de CPTu indicar um avanço na representação estratigráfica de modelos geológicos, a segmentação das camadas com base no comportamento geomecânico apresenta limitações inerentes à subjetividade do processo interpretativo. A análise simultânea de múltiplas variáveis e a atribuição implícita de importância relativa a cada parâmetro depende da experiência e do julgamento do analista. Essa limitação torna-se particularmente relevante em depósitos de natureza complexa e heterogênea, como os solos residuais, onde as transições entre horizontes são frequentemente graduais e a variabilidade espacial é elevada. Nessas condições, a definição dos limites de camada pode variar bastante entre diferentes intérpretes, podendo resultar em inconsistências na classificação do comportamento do solo e na formulação de modelos geotécnicos representativos.

Nesse contexto, técnicas de aprendizado de máquina tornam-se alternativa para automatizar e sistematizar a interpretação de grandes volumes de dados, integrando múltiplos parâmetros em métodos de classificação reproduzíveis. Estudos recentes têm demonstrado o potencial dessas técnicas na classificação automática de solos e na identificação de unidades geotécnicas a partir de dados de CPTu, empregando desde redes neurais artificiais até métodos avançados de *deep learning* (Ghaderi et al., 2019; Xiao et al., 2021; Liu et al., 2024). O presente trabalho apresenta uma prova de conceito para aplicação de aprendizado supervisionado na classificação de horizontes geomecânicos a partir de dados de CPTu. Foram utilizados 32 ensaios realizados em perfil de solo residual, adotando-se condicionais pré-estabelecidas para identificação inicial das camadas correspondentes a Solo Residual Tipo A (mole), Saprólito Tipo A, Solo Residual Tipo B (fofo) e Saprólito Tipo B. Um modelo baseado em Random Forest foi treinado com seis variáveis de entrada (u_0, u_2, q_t, I_b, R_f e B_q), sendo posteriormente aplicado para classificar casos não enquadrados pelas condicionais ou com múltiplas classificações possíveis. Para mitigar o efeito de hipersegmentação característico de classificações pontuais, implementou-se suavização por janela móvel, visando uniformizar transições e definir limites de camada mais representativos da continuidade vertical do maciço.

A abordagem discutida não visa substituir o julgamento do engenheiro geotécnico, mas sim potencializar sua capacidade analítica, transferindo o esforço da interpretação individualizada de dezenas de perfis para a definição de critérios de classificação e validação dos resultados obtidos, com ganhos potenciais em produtividade, repetibilidade e consistência das interpretações.

2 REFERENCIAL TEÓRICO

2.1 CPTu

O ensaio de piezocone (CPTu) constitui um método de investigação in situ baseado na cravação estática instrumentada, permitindo a obtenção simultânea da resistência de ponta corrigida (q_t), do atrito lateral (f_s) e da poropressão (u_2). A aquisição contínua desses parâmetros permite a aplicação de correlações para estimativa de propriedades geotécnicas e fornece subsídios à interpretação do comportamento mecânico, contribuindo para a classificação do solo com base em seu desempenho geomecânico. Conforme descrito por Lunne et al (1997), o uso de ensaios de penetração de cone em campo constitui a técnica mais eficiente para obtenção de dados sobre a estratigrafia e as características físico-mecânicas dos solos em profundidade. O índice de comportamento do solo (I_b) constitui um parâmetro derivado dos resultados do CPTu que integra informações de resistência e atrito, permitindo diferenciar materiais com distintos comportamentos geomecânicos, conforme proposto por Robertson (2016). Sua utilização auxilia na identificação de horizontes e na interpretação de perfis geotécnicos, especialmente em depósitos heterogêneos.

2.2 Random Forest

O Random Forest é um algoritmo de aprendizado supervisionado baseado no princípio de *ensemble learning*, proposto por Breiman (2001). O método combina múltiplas árvores de decisão construídas a partir de amostras aleatórias dos dados, utilizando a técnica de *bootstrap aggregating* (*bagging*). Cada árvore é treinada com um subconjunto amostrado com reposição do conjunto original, enquanto, em cada nó de divisão, apenas um subconjunto aleatório de variáveis é considerado para particionamento.

Formalmente, para um conjunto de treinamento $D = \{(x_i, y_i)\}_{i=1}^{n}$, cada árvore T_bé construída a partir de uma amostra bootstrap D_b. No caso de classificação, a predição final é obtida por votação majoritária e, para regressão, pela média das predições individuais:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b\,(x)$$

Onde B representa o número de árvores na floresta.

Uma das principais propriedades do Random Forest é a redução da variância do estimador, mantendo baixo viés relativo, característica que o torna menos suscetível ao *overfitting* (decorar os dados de treino) quando comparado a árvores individuais.

O método também permite avaliar quais variáveis são mais importantes para a classificação. Uma forma de medir isso é pelo índice de Gini, que calcula o quanto cada variável contribui para separar os dados em grupos mais puros (com menos mistura de classes). Quanto maior a redução da impureza proporcionada por uma variável, mais importante ela é considerada no modelo.

2.3 Janela deslizante

A janela deslizante é um procedimento utilizado na análise de dados sequenciais onde um intervalo de comprimento fixo percorre a extensão dos dados com incrementos regulares. Em cada posição, calcula-se uma medida resumo (média, mediana, moda, etc.) com base nos valores contidos no intervalo. O deslocamento entre posições consecutivas, denominado passo ou *overlap*, determina o nível de suavização e a continuidade entre as análises vizinhas. Valores menores de passo produzem maior sobreposição entre janelas consecutivas, resultando em transições mais graduais entre os segmentos. A técnica é aplicada quando se deseja reduzir flutuações locais e extrair padrões predominantes ao longo de uma sequência, agrupando observações adjacentes com comportamentos semelhantes (Datar et al., 2002). Em termos geotécnicos, a janela deslizante atua como um filtro para suavizar as oscilações pontuais observadas nos perfis obtidos pelo ensaio de cone (CPTu) e nas propriedades derivadas por correlações empíricas. Ao agrupar valores consecutivos em intervalos móveis, a técnica permite identificar tendências predominantes do solo.

3 MATERIAL E MÉTODOS

3.1 Ensaios realizados

Foram utilizados 32 ensaios CPTu realizados em perfil de alteração contendo solos residuais e saprolitos de dois litotipos distintos, adotando-se condicionais pré-estabelecidas para identificação inicial das camadas correspondentes a cada unidade geotécnica. Essas classificações foram definidas com base na análise geológica local, considerando o contexto litológico, os processos de intemperismo predominantes e o modelo geológico-geotécnico previamente estabelecido para a área de estudo.

A interpretação dos ensaios CPTU para definição das unidades geológico-geotécnicas foi fundamentada no comportamento predominante do material ao longo do perfil, distinguindo-se domínios do tipo *sand-like* (arenoso) e *clay-like* (argiloso). Essa distinção foi realizada a partir da análise integrada dos principais parâmetros obtidos no ensaio: índice de comportamento do solo (I_b), resistência de ponta corrigida (q_t),

parâmetro de poropressão (B_q), razão de atrito (F_r) e a relação entre a poropressão medida (u_2) e a poropressão estática (u_0). A Tabela 1 apresenta os intervalos limites considerados por parâmetro do ensaio CPTu.

Tabela 1 – Limites dos parâmetros do ensaio CPTu considerados para classificação dos solos.

Unidade	Grupo	Domínio	I_b	q_t (MPa)	B_q	Relação u_2/u_0	F_r (%)
Solo Residual Tipo A	2	Clay-like	< 25	< 3	> 0,2	$u_2 > u_0$	> 2,5
Saprólito Tipo A	4	Clay-like	< 32	< 10	-0,1 < Bq < 0,2	$u_2 < u_0$	> 2,0
Solo Residual Tipo B	1	Sand-like	22 < Ib < 32	< 5	< 0,2	–	1 < FR < 3,5
Saprólito Tipo B	3	Sand-like	> 32	> 5	< 0,1	$u_2 < u_0$	1 < FR < 3

Os materiais enquadrados como *sand-like* apresentam características associadas a comportamento predominantemente drenado, menor desenvolvimento de excesso de poropressão e resposta mecânica típica de solos granulares ou silto-arenosos. Já os materiais classificados como *clay-like* evidenciam maior influência da geração de poropressão durante a cravação, valores relativos de resistência de ponta mais baixos e comportamento compatível com solos de matriz fina.

Destaca-se que os limites apresentados representam faixas típicas de comportamento e não fronteiras rígidas entre materiais, uma vez que solos residuais frequentemente exibem transições graduais entre domínios mais arenosos e mais argilosos. Dessa forma, o enquadramento foi conduzido de maneira interpretativa e iterativa, buscando coerência entre os parâmetros medidos no CPTU e o comportamento geotécnico esperado para cada unidade.

3.2 Processamento dos métodos

A metodologia adotada integra uma etapa de classificação determinística, baseada nos intervalos limites pré-definidos dos parâmetros, seguida de um modelo supervisionado de aprendizado de máquina, visando reduzir ambiguidades na definição de horizontes geomecânicos a partir de dados de CPTu. Inicialmente, cada registro do ensaio é submetido a uma rotina de classificação por regras condicionais previamente estabelecidas (Tabela 1). São consideradas seis variáveis principais: poropressão medida no cone (u_2), resistência de ponta corrigida (q_t), razão de atrito (F_r), índice de comportamento do solo (I_b), parâmetro de poropressão (B_q) e poropressão estática (u_0). A partir de combinações específicas dessas variáveis, são classificados quatro grupos geotécnicos distintos (Grupo 1 a Grupo 4). Quando todas as condições são satisfeitas, o ponto é classificado diretamente no respectivo grupo. Caso nenhuma condição seja atendida, o registro recebe o rótulo "Não Classificado"; quando o ponto se enquadra em mais de um grupo de condição, é marcado como "Múltiplos Grupos", indicando incerteza na interpretação determinística.

Os registros classificados de forma inequívoca (Grupo 1 a 4) constituem o conjunto de treinamento do modelo supervisionado. As variáveis de entrada (u_2, q_t, F_r, I_b, B_q e u_0) são previamente padronizadas por meio de normalização (*standardscaler*), assegurando que diferenças de escala não influenciem o desempenho do algoritmo. Foram utilizados *n_estimators* = 100 e *max_depth* = 4. Segundo Probst, Wright e Boulesteix (2019), o aumento do número de árvores melhora a estabilidade do modelo, mas os ganhos tendem a diminuir após certo ponto, justificando o uso de 100 árvores. Além disso, a limitação da profundidade para quatro níveis reduz a complexidade do modelo e o risco de sobreajuste, sendo a profundidade um hiperparâmetro que deve ser ajustado conforme as características dos dados. O conjunto de dados é dividido em subconjuntos de treino e teste (80%/20%), permitindo avaliar o desempenho do modelo.

Após o treinamento, o modelo é aplicado exclusivamente aos registros inicialmente rotulados como "Não Classificado" ou "Múltiplos Grupos". Para esses casos, o Random Forest fornece tanto a classe prevista quanto as probabilidades associadas a cada grupo possível. A classificação final passa a considerar explicitamente o nível de confiança da predição: quando a probabilidade do grupo estimado é superior a 60%, a classificação é aceita como válida; quando igual ou inferior a esse limiar, o resultado é mantido, porém identificado como "baixa confiança".

Dessa forma, a metodologia estabelece três níveis de decisão:

(i) classificação original determinística, quando inequívoca;
(ii) classificação por aprendizado de máquina independentemente da confiança para a classe;
(iii) classificação devido à confiança da classificação em cada classe.

Após a etapa de classificação, aplica-se um procedimento de suavização por janela deslizante ao longo do perfil de profundidade, visando reduzir as segmentações excessivas comuns em dados de CPTu. Para isso, utiliza-se uma janela de 1 m de espessura, deslocada com sobreposição de 50% (*overlap* = 0,5) entre janelas consecutivas. Em cada intervalo, determina-se o grupo geotécnico predominante, permitindo obter um perfil estratigráfico mais contínuo e representativo das tendências do solo.

4 RESULTADOS

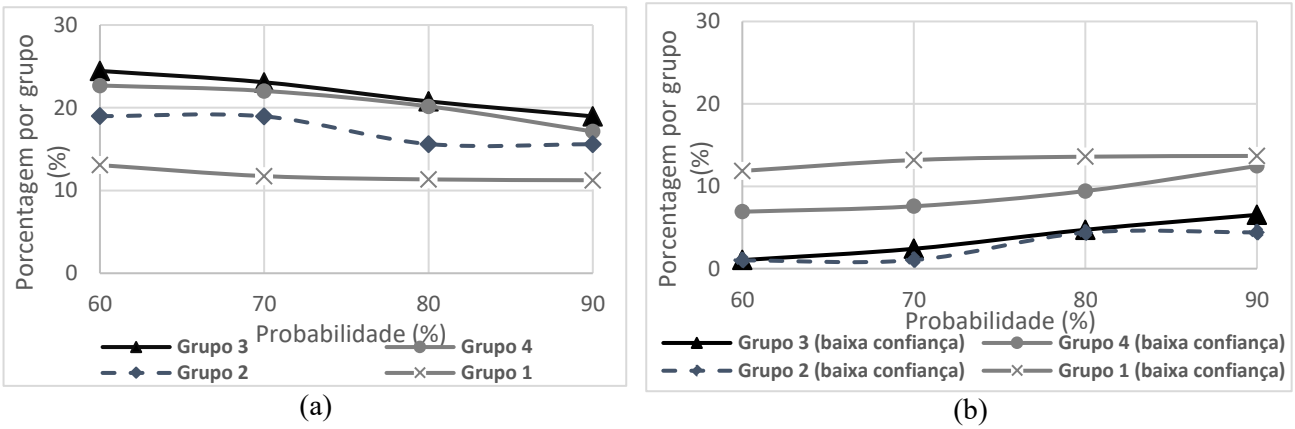
Segundo os resultados obtidos, observa-se que, na etapa de classificação determinística baseada em condicionantes, a maior parte dos registros, 6613 pontos (40,81%), não atendeu a nenhuma das condições definidas, sendo inicialmente rotulados como "Não Classificado". Um pequeno percentual, 648 pontos (4,00%), apresentou condições sobrepostas, sendo categorizado como "Múltiplos Grupos", ou seja, estes pontos se encaixaram em mais de um grupo classificado. Na maior parte dos casos, a sobreposição ocorreu entre os Grupos 1 e 4, indicando incertezas nos limites de classificação entre essas camadas. Entre os pontos classificados de forma inequívoca, o Grupo 4 apresentou a maior quantidade, com 2752 registros (16,98%), seguido pelo Grupo 2 com 2237 registros (13,82%), Grupo 3 com 2136 registros (13,18%) e Grupo 1 com 1817 registros (11,21%). Portanto, a distribuição dos grupos é satisfatória, com variedade e representatividade adequadas para o treinamento do modelo. Os registros inequívocos correspondem a 55,19% do conjunto total, oferecendo uma base consistente para análises subsequentes e para o treinamento de modelos de aprendizado de máquina. A distribuição dos registros mostra que uma parte significativa dos pontos (40,81%) não se enquadra nos grupos previamente definidos, e uma parcela menor (4,00%) atende simultaneamente a mais de um critério. Esses resultados indicam limitações da classificação determinística condicional.

Os registros inicialmente rotulados como "Não Classificados" foram posteriormente processados por um modelo Random Forest, treinado com os registros inequívocos (Grupos 1 a 4) da etapa condicional determinística. Para verificar a presença de overfitting, o desempenho do modelo foi avaliado nos conjuntos de treino e teste. Observou-se acurácia de 0,999 no treino e 0,998 no teste, indicando excelente capacidade de generalização e mínima diferença entre os conjuntos, o que sugere que o modelo aprendeu os padrões dos dados determinísticos sem memorizar os dados de treinamento.

A importância das variáveis foi avaliada pelo critério Gini importance. O parâmetro Bq apresentou a maior contribuição para a classificação, seguido por Ib. As variáveis Fr, u_2 e qt tiveram influência intermediária, enquanto u_0 apresentou contribuição desprezível para o desempenho do modelo.

A aplicação do modelo Random Forest aos registros inicialmente problemáticos permitiu aumentar a quantidade de pontos atribuídos a grupos geotécnicos em relação à classificação determinística. O Random Forest não fornece apenas uma previsão de grupo, mas também a probabilidade associada a cada classe (*predict_proba*), refletindo o grau de confiança do modelo em sua decisão. Essa informação probabilística foi utilizada para estabelecer diferentes limiares de confiança (60%, 70%, 80% e 90%) permitindo distinguir predições consideradas confiáveis daquelas de baixa confiança e, assim, fornecer uma interpretação mais segura do perfil geotécnico.

Com o limiar de 60%, 79,2% dos registros foram classificados com confiança adequada nos grupos de maior confiabilidade, enquanto 20,8% foram identificados como de baixa confiança, ou seja, um aumento de 24,01% em relação à classificação determinística. À medida que o limiar aumenta, o número de registros classificados com confiança diminui, refletindo critérios mais rigorosos: com 90% de probabilidade, 64,14% dos registros permanecem em grupos de maior confiança, um aumento de apenas 8,95%. A parcela de baixa confiança cresce proporcionalmente. Assim, considerando que o limiar de 60% representa um equilíbrio adequado entre nível de confiança e a porcentagem de registros classificados, esse valor foi adotado como critério para a aplicação do método de janelas deslizantes. A Figura 1(a) apresenta a porcentagem de registros classificados com maior confiança por grupo em função do limiar de confiança adotado, enquanto a Figura 1(b) mostra a porcentagem de registros classificados com baixa confiança por grupo em função desse mesmo limiar.



(a) (b)

Figura 1 - Variação da porcentagem de classificações por grupo em função do limiar de confiança adotado no modelo Random Forest: (a) registros classificados com maior confiança; (b) registros classificados com baixa confiança.

Com a aplicação do procedimento de janela deslizante, o perfil originalmente caracterizado por uma segmentação excessiva foi suavizado, resultando em uma representação estratigráfica mais contínua, como pode ser observado no exemplo do ensaio da Figura 2. A classificação de cada intervalo passou a considerar o grupo predominante entre os pontos contidos na janela. Apesar da melhoria na continuidade do perfil, algumas limitações foram observadas na metodologia. Em especial, verificou-se que a direção de varredura da janela ao longo da profundidade pode influenciar a definição dos contatos entre camadas, uma vez que o conjunto de pontos considerados em cada posição da janela não é perfeitamente simétrico em relação às transições litológicas. Esse aspecto indica que o método ainda pode ser aperfeiçoado, especialmente no que se refere ao critério de agrupamente das janelas e à representação das interfaces entre materiais distintos.

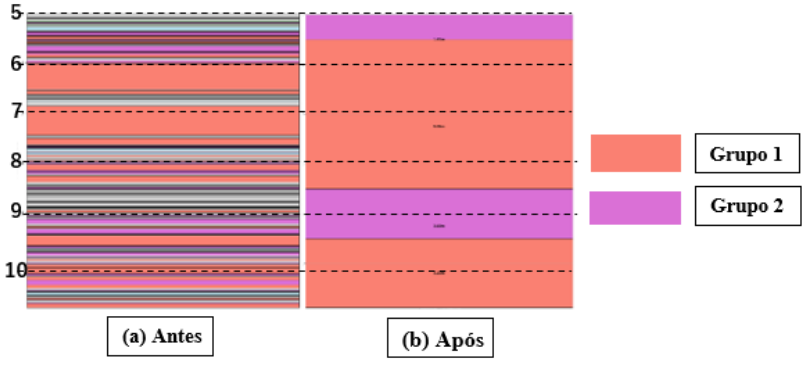


Figura 2 - Perfil geotécnico obtido antes e após a aplicação do método de janela deslizante.

No que tange a comparação entre metodologia de acordo com a porcentagem por grupos, a análise humana apresenta predominância do Grupo 2 (47,25%) e menor representatividade do Grupo 4 (10,86%), enquanto a abordagem apenas determinística deixa 40,81% do perfil não classificado, evidenciando limitações na definição de zonas de transição. Com a aplicação do Random Forest sem adoção do critério de confiança, todo o perfil passa a ser classificado, com redistribuição mais equilibrada entre os grupos e aumento expressivo do Grupo 4 (30,82%), indicando sensibilidade do modelo a padrões que podem não ter sido plenamente capturados na interpretação manual.

Ao aplicar o critério de confiança de 60%, parte dos dados (20,81%) é sinalizada como de baixa confiabilidade, introduzindo um controle objetivo de incerteza sem comprometer significativamente a cobertura. A suavização por janela, por sua vez, promove maior coerência estratigráfica e mantém o Grupo 4 em patamar substancialmente superior ao da análise humana (26,88%), reforçando a hipótese de que este grupo pode ter sido subestimado na interpretação inicial. Assim, os resultados da prova de conceito indicam que a combinação entre aprendizado de máquina, análise de confiança e pós-processamento é tecnicamente viável e potencialmente mais sensível à identificação de determinadas camadas, especialmente aquelas associadas ao Grupo 4. Tabela 2 mostra a distribuição dos registros por grupo e por metodologia de classificação aplicada.

Tabela 2 - Comparação da distribuição dos registros por grupo e por metodologia de classificação.

Classe	Análise humana	Análise determinística	RF sem análise de confiança	RF com análise de confiança (60%)	RF com suavização por janela
Grupo 1	17,56	11,21	23,79	13,07	18,31
Grupo 2	47,25	13,82	20,68	18,99	29,47
Grupo 3	24,33	13,18	24,72	24,45	25,34
Grupo 4	10,86	16,98	30,82	22,69	26,88
Não classificado	-	40,81	-	-	-
Baixa confiança	-	-	-	20,81	-
Múltipla classificação	-	4,00	-	-	-

Comparando metodologias que resultaram na classificação completa dos registros, é possível inferir que o Grupo 4 (Saprólito Tipo A) apresentou subnotificação na análise humana inicial. Esse comportamento pode estar associado a fatores como maior dispersão do grupo ao longo dos perfis ou presença de camadas de pequena espessura, que dificultam sua identificação direta na interpretação manual dos ensaios. Por outro lado, o Grupo 3 (Saprólito Tipo B) manteve participação relativamente constante entre as diferentes metodologias, o que pode indicar que as características associadas a esse material são bem definidas e agrupadas nos ensaios CPTu, permitindo sua identificação de forma consistente tanto na análise humana quanto nas abordagens automatizadas. O Grupo 2 (Solo Residual Tipo A) apresentou variação mais significativa, com predominância na interpretação humana e redução relativa nas classificações automatizadas, sugerindo que parte dos registros inicialmente atribuídos a esse grupo na análise humana pode ter sido classificada de forma excessiva possivelmente devido à contextualização geológica. Já o Grupo 1 (Solo Residual Tipo B) também apresentou variações entre as metodologias, porém em menor magnitude, indicando relativa estabilidade na sua identificação ao longo das diferentes abordagens.

5 CONCLUSÕES

A prova de conceito apresentada demonstra que a aplicação de algoritmos de aprendizado de máquina, em particular o Random Forest, pode contribuir para aprimorar a interpretação de ensaios CPTu. O modelo alcançou acurácia de 0,998 no conjunto de teste, indicando elevada capacidade de reconhecer padrões presentes nas regras condicionais determinísticas dos parâmetros analisados. Como resultado, houve redução significativa dos registros não classificados pelas regras iniciais, que deixavam mais de 40% dos dados sem uma definição clara. A análise de importância das variáveis indicou que os parâmetros B_q e I_b exercem maior

influência na distinção entre os grupos identificados, sugerindo que esses índices desempenham papel mais relevante na caracterização dos horizontes investigados. Além disso, a introdução de critérios de confiança estatística associada à aplicação de uma suavização por janela móvel contribuiu para reduzir a fragmentação excessiva dos perfis, resultando em modelos estratigráficos mais contínuos e compatíveis com a interpretação geológica esperada. Os resultados também indicaram que a abordagem automatizada apresentou com maior frequência a identificação de determinadas camadas, como aquelas associadas ao Grupo 4, que foram menos presentes em interpretações manuais tradicionais. Observou-se ainda que existem variações na participação relativa dos grupos entre as diferentes metodologias de classificação, o que pode indicar possíveis sub ou super-representações de determinadas classes quando a análise depende exclusivamente da interpretação humana. Dentre as principais limitações identificadas no método proposto destacam-se a circularidade metodológica inerente ao processo de treinamento, o efeito do sentido de varredura da janela móvel de suavização e a desconsideração explícita dos perfis de intemperismo dos materiais. Como o modelo foi treinado com classes definidas pelas próprias regras condicionais utilizadas na classificação inicial, sua acurácia deve ser interpretada como a capacidade de reproduzir esses critérios, e não como uma validação independente do comportamento geomecânico dos materiais. Todas elas são passíveis de aprimoramento em evoluções futuras. Dessa forma, a metodologia proposta não tem como objetivo substituir o julgamento do engenheiro geotécnico, mas atuar como uma ferramenta de apoio à análise dos dados. Ainda assim, os resultados apresentados devem ser interpretados como uma etapa inicial de investigação, sendo recomendável a realização de estudos adicionais, a fim de avaliar a coerência e a representatividade das camadas interpretadas perante o modelo geológico local e o perfil de intemperismo dos domínios litológicos.

REFERÊNCIAS BIBLIOGRÁFICAS

Breiman, L. (2001) *Random forests. Machine Learning*, Dordrecht, v. 45, n. 1, p. 5-32, out. https://doi.org/10.1023/A:1010933404324.

Datar M, Gionis A, Indyk P, Motwani R (2002) *Maintaining stream statistics over sliding windows*. SIAM J Comput 31(6):1794–1813windows. SIAM J Comput 31(6):1794–1813

Ghaderi, A.; Abbaszadeh Shahri, A.; Larsson, S. (2019) *An artificial neural network based model to predict spatial soil type distribution using piezocone penetration test data.* Bulletin of Engineering Geology and the Environment, v. 78, n. 6, p. 4579–4588. https://doi.org/10.1007/s10064-018-1400-9.

Liu, W.; Tong, L.; Yan, X.; Liu, S. et al. (2024) *Automatic soil classification method from CPTU data based on convolutional neural networks*. Bulletin of Engineering Geology and the Environment, v. 83, n. 8. https://doi.org/10.1007/s10064-024-03815-6.

Lunne, T.; Robertson, P. K.; Powell, J. J. M. (1997) *Cone penetration testing in geotechnical practice*. Londres: Blackie Academic & Professional, 1997.

Probst P, Wright, M. N, Boulesteix, A. L (2019) *Hyperparameters and tuning strategies for random forest*. WIREs Data Mining and Knowledge Discovery, v. 9, n. 3, e1301.

Robertson, P. K. (2016) *Cone Penetration Test (CPT)-Based Soil Behaviour Type (SBT) Classification System—An Update.* Canadian Geotechnical Journal, 53, 1910-1927. https://doi.org/10.1139/cgj-2016-0044

Xiao, T.; Du, Y.; Zou, H. F.; Yin, K. S.; Zhang, L. M. (2021) *Machine learning-enhanced soil classification by integrating borehole and CPTU data with noise filtering.* Bulletin of Engineering Geology and the Environment, v. 80, n. 12, p. 9157–9171. https://doi.org/10.1007/s10064-021-02478-x.