



OTIMIZAÇÃO DE HIPERPARÂMETROS COM ALGORITMO GENÉTICO: UMA ANÁLISE COMPARATIVA ENTRE ANN E RF PARA PREDIÇÃO DO ÍNDICE DE SUPORTE CALIFÓRNIA

Iago de Souza Lopes¹, Rogério Ramos Moreira¹, Alexandre Ribeiro¹, Marta Pereira da Luz¹

¹Pontifícia Universidade Católica de Goiás

RESUMO

O tempo demandado para realizar o ensaio do Índice de Suporte Califórnia (ISC) justifica a busca por métodos preditivos computacionais. No entanto, modelos de Machine Learning frequentemente carecem de uma metodologia de validação criteriosa, arriscando conclusões superestimadas. Visando contribuir com esta temática, este trabalho propõe um framework robusto para a otimização e comparação de modelos ANN e RF, via Algoritmo Genético (GA). A metodologia avalia sistematicamente múltiplos cenários experimentais a partir de banco de dados público da COPPE/UFRJ. Os resultados demonstraram que a abordagem de otimização é eficaz, com o modelo GA-ANN, alcançando um R^2 de 0.780 nos dados de teste. O framework apresentado consiste em uma ferramenta eficaz para o desenvolvimento de modelos preditivos, agilizando a estimativa do CBR para dar suporte às análises preliminares e decisões de anteprojeto em pavimentação.

1. INTRODUÇÃO

A engenharia geotécnica é fundamental para a estabilidade de projetos de infraestrutura, como as rodovias, que impactam diretamente a economia e a logística de um país. No Brasil, o método de dimensionamento de pavimentos flexíveis depende fortemente do ensaio do Índice de Suporte Califórnia ou *Califórnia Bearing Ratio* (CBR), um processo laboratorial demorado e caro, representando até 71% do tempo total de identificação de uma amostra de solo (YABI et al., 2024).

Em resposta a esses desafios, pesquisadores têm aplicado algoritmos de Machine Learning (ML), como Artificial Neural Networks (ANN) e Random Forest (RF), para estimar o CBR, modelando as complexas relações não lineares do comportamento do solo (TADO et al., 2025). Contudo, o desempenho desses modelos de ML é altamente sensível à escolha de seus hiperparâmetros. Técnicas de otimização, como os Algoritmos Genéticos (GA) e Algoritmos de Enxame de Partículas (PSO), podem automatizar essa busca possibilitando a diminuição de gastos, em comparação à métodos de tentativa e erro. Como em estudos feitos por Bardhan et al. (2021) e por Yabi et al. (2024).

Para automatizar essa busca, este trabalho emprega a técnica do Algoritmo Genético (GA), que pertence a uma classe de métodos de otimização conhecidos como algoritmos evolucionários. Ele é inspirado em processos da natureza como a seleção natural. Em vez de refinar uma única solução, o GA trabalha com uma população de "indivíduos", neste caso cada um possui diferentes configurações de hiperparâmetros, que estão espalhados pelo espaço de busca. Essa abordagem populacional permite que o algoritmo realize uma busca inerentemente global, reduzindo a tendência de ficar preso em ótimos locais. Adicionalmente, possui processos como o cruzamento (*crossover*) e a mutação que são utilizados para explorar novas regiões do espaço de busca. Contudo, como uma meta-heurística, o GA não garante matematicamente a convergência para o ótimo global, mas aumenta significativamente a probabilidade de encontrar soluções de melhores em problemas complexos e não-lineares, como a otimização de hiperparâmetros (MARTINS e NING, 2021).

Além disso, apesar do uso de ML na área, muitos trabalhos não empregam uma metodologia de validação robusta, negligenciando uma separação mais criteriosa entre dados de treino, validação e teste, ou a avaliação da variabilidade estocástica dos modelos, o que pode levar a conclusões de desempenho superestimadas devido ao *overfitting*.

Diante deste cenário, este trabalho propõe um *framework* que visa, além de otimizar os hiperparâmetros de ANN e RF via GA, possibilita também a avaliação dos modelos de maneira mais abrangente, através de técnicas de *Cross-Validation*, *Selection Features* e uso de pesos em algumas amostras. Assim, o objetivo central do estudo é analisar como a otimização de hiperparâmetros, realizada pelo Algoritmo Genético afeta o desempenho comparativo dos modelos ANN e RF na predição do CBR, quando avaliados sob uma metodologia robusta.

2. METODOLOGIA

O estudo utilizou um banco de dados público da COPPE/UFRJ (FERREIRA, 2008), originalmente com 463 amostras de solos. O tratamento dos dados envolveu a remoção de colunas não pertinentes à predição de CBR. Sendo que, as colunas de interesse referem-se a parâmetros geotécnicos de caracterização dos solos, como limites de Atterberg (LL, IP) e dados de granulometria (D1-D6), além a exclusão de todas as amostras com valores ausentes em qualquer uma das colunas de interesse, resultando em um conjunto final de cerca de 330 amostras dependendo da seleção das features de input, definidas no *feature selection* dependendo da configuração experimental definida pelo usuário da ferramenta.

A análise exploratória revelou um forte desequilíbrio na variável-alvo (CBR), em que 75% das amostras apresentaram valores de CBR inferiores a 40 (%). O objetivo do modelo é fazer previsões precisas em toda a gama de valores, contudo essa concentração em valores baixos pode distorcer o treinamento, fazendo com que o modelo negligencie valores altos, que são menos presentes no dataset. Para mitigar esse viés, este estudo testa uma estratégia de ponderação atribuindo maior peso durante o treinamento às amostras de CBR alto (maiores do que 40%) que já existem na base de dados (GOOGLE DEVELOPERS, s.d.). O objetivo alvo é fazer com que o modelo ponderado demonstre uma capacidade preditiva mais equilibrada, embora ainda sujeita a limitações, como a subestimação de valores de CBR muito altos.

O conjunto de desenvolvimento é então utilizado no processo de otimização, onde um Algoritmo Genético (GA) busca a melhor combinação de hiperparâmetros para o modelo em questão (ANN ou RF). A aptidão (*fitness*) de cada configuração candidata durante a otimização é avaliada utilizando um dos dois métodos de validação: *Holdout* (divisão interna do conjunto de desenvolvimento) ou *K-Fold*. Para garantir a estabilidade contra a variabilidade estocástica do treinamento, a avaliação de cada indivíduo é repetida múltiplas vezes, e a média dos resultados é utilizada como a pontuação final de *fitness*.

Ao final do processo de otimização, a melhor configuração encontrada é utilizada para treinar um modelo final, cuja performance de generalização é medida no conjunto de teste.

2.1. Configuração experimental

Para avaliar os modelos de forma abrangente foi definida uma matriz de cenários experimentais, variando sistematicamente os seguintes fatores:

- Seleção de Modelo: Otimização para GA-ANN e GA-RF.
- Seleção de Features: Testes com alguns subconjuntos de features: D1: [D6, CH, CY]; D2: [D6, CH, CY, IP]; D3: [D6, CH, CY, IP, LL]; D4: [IG, EXP, D3, D4, D5, D6, CH, CY, IP, LL].
- Tratamento de Desbalanceamento: Execuções com e sem a aplicação de pesos para amostras de CBR alto.
- Método de Validação: Comparação entre *Holdout* e *K-Fold* (com K=5).

2.2. Modelagem e otimização

- Otimização: O GA foi utilizado para encontrar a melhor combinação de hiperparâmetros. A aptidão de cada indivíduo é calculada como a média da performance em múltiplas repetições para reduzir a variação estocástica dos modelos, ou seja, buscar o modelo mais estável.
- Avaliação Robusta: Após encontrar a melhor configuração, o modelo final é treinado e avaliado 10 vezes com diferentes sementes aleatórias. As métricas de performance (R^2 e MAE) são reportadas como a média.
- Análise: Os resultados são salvos em um diretório único, permitindo análises de convergência, viés e importância de features. A reprodutibilidade é garantida pelo uso de uma semente aleatória fixa e pela disponibilização do código-fonte presente no link do projeto: github.com/iagoLopex/TCC.

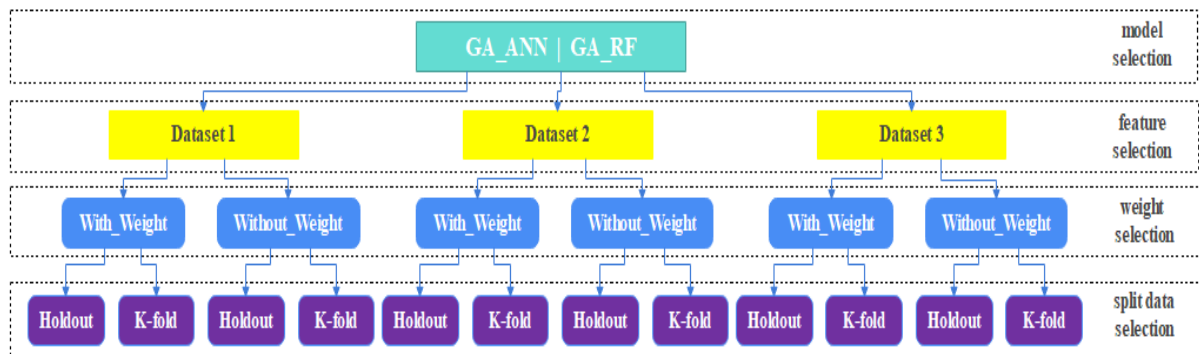


Figura 1: Fluxograma do Framework

3. RESULTADOS E DISCUSSÃO

3.1. Análise comparativa de desempenho

A análise da Tabela 1 revela que, ao utilizar o conjunto de features mais completo (D4) e o método de validação Holdout, o modelo GA-ANN alcançou o maior R^2 no conjunto de teste (0.784) e um GAP de +0.007, demonstrando a performance preditiva mais elevada entre os cenários testados e o desenvolvimento de uma melhor generalização tanto das amostras de validação quanto de teste, pelo baixo valor do GAP, além validar que neste caso as redes neurais são melhores para capturar interações complexas entre muitas características. O modelo GA-RF, embora mais estável nos datasets mais simples (D1-D3), atingiu um pico de performance inferior ($R^2 = 0.738$) nas mesmas condições e um GAP de +0.159 (o modelo conseguiu obter acurácia muito melhor com as amostras de validação do que com as amostras de teste). Isso sugere que a arquitetura flexível das redes neurais, quando otimizada pelo GA, foi mais capaz de extrair padrões preditivos complexos a partir do conjunto de dados com mais informações de *input*.

Tabela 1: Resultados dos modelos de predição do CBR

Modelo	Entrada	Método de Validação	R^2 Validação	R^2 Teste	GAP (Val-Test)
GA-RF	D1	Holdout	0.783	0.718	+0.065
	D2	Holdout	0.786	0.716	+0.070
	D3	Holdout	0.789	0.721	+0.068
	D4	Holdout	0.897	0.738	+0.159
	D4	K-fold	0.753	0.751	+0.002
GA-ANN	D1	Holdout	0.816	0.677	+0.139
	D2	Holdout	0.803	0.694	+0.109
	D3	Holdout	0.852	0.616	+0.236
	D4	Holdout	0.791	0.784	+0.007
	D4	K-fold	0.798	0.756	+0.042

3.2. Análise de overfitting e vies do modelo

A Figura 2 demonstra a capacidade do melhor modelo em aprender a tendência geral dos dados de CBR (%), com a linha de predição próxima aos valores reais, especialmente para CBRs menores, capturando suas variações. Contudo, essa visualização revela uma limitação crítica: o modelo subestima picos de alto valor e superestima outros valores, falhando em alcançar a magnitude dos valores verdadeiros para as amostras mais elevadas (e.g., índices 3, 14, 24). Este comportamento é consistente com o desbalanceamento do dataset, onde a escassez de amostras com CBR elevado dificulta o aprendizado do modelo nessas faixas, reforçando a necessidade de análises de vies e estratégias de tratamento de dados para futuros aprimoramentos.

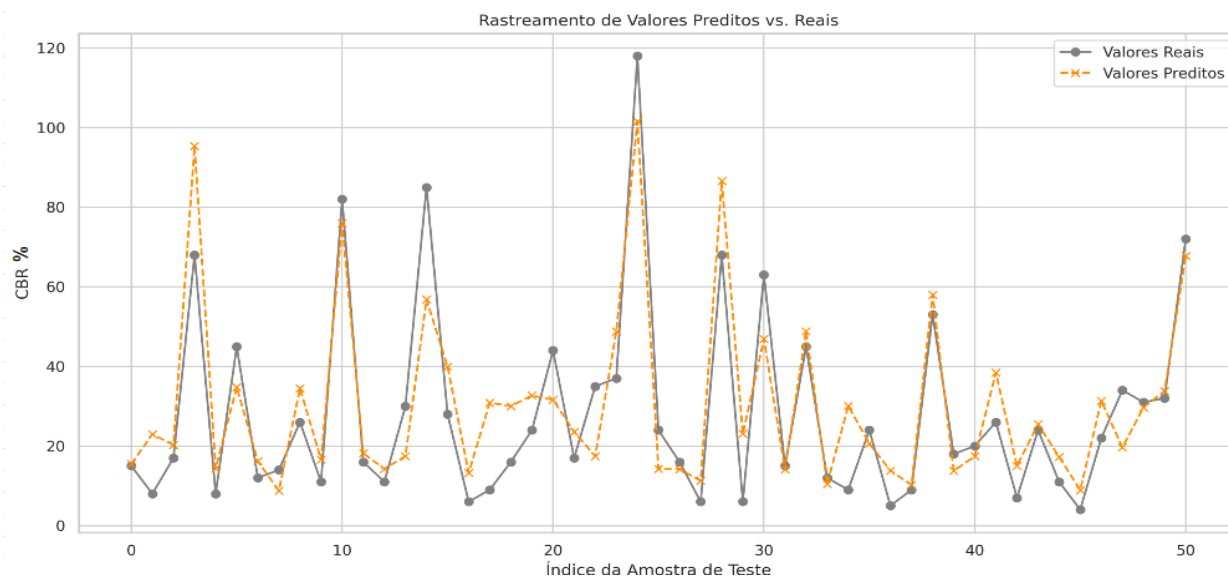


Figura 2: Gráfico Real vs. Predito do melhor modelo

4. CONCLUSÃO

Este trabalho demonstrou que a otimização de hiperparâmetros por Algoritmos Genéticos aprimora significativamente modelos de Machine Learning para predição de CBR, com o GA-ANN, aliado ao conjunto de features mais completo (D4), alcançando o R^2 de 0.784 no teste. Além disso, análise comparativa evidenciou que a arquitetura flexível da ANN foi mais capaz de extrair padrões do conjunto de dados expandido, enquanto o GA-RF, apesar da estabilidade, apresentou maior dificuldade de generalização com o mesmo conjunto de features, indicado por um GAP (Val-Teste) superior.

Outro fator foi que a análise de viés do modelo final revelou uma limitação importante, em que a tendência de subestimar picos de alto valor de CBR, é um fator de comportamento consistente com o desbalanceamento identificado no dataset. Futuramente, portanto, sugere-se como passo principal a aplicação de técnicas de reamostragem para mitigar o viés na predição de valores elevados. Adicionalmente, recomenda-se a exploração de outros algoritmos, como o Gradient Boosting, para expandir os resultados obtidos e aplicações novas técnicas de metaheurísticas.

REFERÊNCIAS BIBLIOGRÁFICAS

- BARDHAN, A.; P. Samui; K. Ghosh; A. H. Gandomi e S. Bhattacharyya (2021) ELM-based adaptive neuro swarm intelligence techniques for predicting the California bearing ratio of soils in soaked conditions. *Applied Soft Computing*, v. 110, p. 107595. DOI: 10.1016/j.asoc.2021.107595.
- FERREIRA, J. G. H. M. Tratamento de Dados Geotécnicos Para Predição de Módulos de Resiliência de Solos e Britas Utilizando Ferramentas de Data Mining. 2008. Tese (D.Sc. em Engenharia Civil) - COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2008.
- Google Developers. (s.d.). Conjuntos de dados desequilibrados. Machine Learning Crash Course. Disponível em: <https://developers.google.com/machine-learning/crash-course/classification/handling-imbalanced-classes>. Acesso em: 5 jun. 2025.
- MARTINS, J. R. R. A. e A. Ning (2021) *Engineering Design Optimization*. Cambridge: Cambridge University Press.
- TADO, N.; MEDIHAJIT, S.; PAL, D. Forecasting California bearing ratio (CBR) of soil using machine learning algorithms: A review. *Research on Engineering Structures and Materials*, v. 11, n. 1, p. 383-398, 2025.
- YABI, C. P. et al. Prediction of CBR by Deep Artificial Neural Networks with Hyperparameter Optimization by Simulated Annealing. *Indian Geotechnical Journal*, v. 54, n. 1, p. 121-137, fev. 2024.