

**USANDO APRENDIZADO DE MÁQUINA EM ANÁLISES PREDITIVAS
SOBRE PEQUENOS DATASETS DE DADOS DE SOLOS**

Caroline Silva De Oliveira (caroline@pet-si.ufrj.br)

Sérgio Manuel Serra Da Cruz (serra@pet-si.ufrj.br)

O solo é um recurso natural, limitado e essencial para o desenvolvimento da vida. Visando aproveitar de modo mais sustentável tais recursos, este estudo propõe uma abordagem computacional, que consiste na utilização de um dataset com centenas de registros, para a identificação dos tipos de solo, utilizando técnicas de Inteligência Artificial a partir da plataforma OpenSoils (www.opensoils.org) que reúne dados abertos de solos brasileiros (Marinho et al., 2023). A plataforma está passando por um processo de modernização para se tornar mais ágil e assertiva na análise de dados. OpenSoils já segue os padrões de Open Science e e-Science sugeridos pela comunidade científica internacional, assim como os Princípios FAIR para o gerenciamento de grandes volumes de dados. Como metodologia de pesquisa, foi escolhida a técnica de aprendizado de máquina supervisionado denominada árvore de decisão, recomendada por suas regras simples no tratamento de dados de baixa complexidade. Foram escolhidos os dados referentes a utilização do solo no perímetro da Fazendinha Agroecológica km 47, Seropédica – RJ. Pois trata-se de um tradicional espaço de pesquisas em agricultura orgânica, mantido por uma parceria entre a Embrapa, a Pesagro-Rio e a UFRRJ. No entanto, apesar de sua relevância e importância, a Fazendinha possui pequena quantidade de perfis de solos, consequentemente, poucos dados de solos. Os códigos da

contribuição computacional utilizaram apenas máquinas de baixíssimo custo com poucos recursos computacionais e foram produzidos em linguagem Python diretamente nas interfaces do Google Colab. Este código extrai os dados de um dataset maior e realiza as etapas de tratamento de dados do ciclo de vida da engenharia de dados, como, por exemplo, a limpeza, o preenchimento de dados faltantes, o tratamento de dados outliers para evitar o overfitting e, finalmente, separa o arquivo em duas partes (até 60% dos dados para treinamento e até 40% para testes). Os dados referem-se aos tipos de solos encontrada na área. Visando várias análises de desempenho e assertividade sobre pequenos datasets, variou-se o quantitativo de dados de teste entre 10% e 40%, com intervalos de 5% entre as amostras, para compreender a relação entre as métricas de acurácia, precisão, recall e F1-Score em relação ao tamanho do dataset. A acurácia indica a quantidade de acertos do nosso modelo dividido pelo total de dados. A precisão avalia os verdadeiros positivos, ou seja, a quantidade de vezes que o modelo acerta em relação ao total de vezes que ele tenta acertar. O recall avalia os verdadeiros negativos, quantificando as vezes que o modelo acerta em relação ao total de vezes que ele deveria ter acertado. Já o F1-Score é uma métrica que combina precisão e recall de maneira equilibrada. Os resultados dos experimentos iniciais mostram os seguintes valores: Para 20% dos dados: acuraria = 0,364; Precisão = 0,438; Recall = 0,5 e F1-Score = 0,458; para 25% dos dados: acuraria = 0,538; Precisão = 0,494; Recall = 0,625 e F1-Score = 0,531 e, para 30% dos dados: acuraria = 0,438; Precisão = 0,472; Recall = 0,625 e F1-Score = 0,5. Conclui-se que, tendo o F1-Score como métrica base, o valor de 25% sobre os dados de teste apresenta o melhor desempenho em pequenos datasets de dados. Os experimentos mostram que, mesmo diante de amostras de tamanho limitado, ainda é possível aplicar classificadores de dados; no entanto, a baixa eficiência (em torno de 53%) em datasets com poucos registros indica que novos métodos analíticos podem ser explorados.

REFERÊNCIAS:

MARINHO, É. C. et al. OpenSoils: Uma Plataforma de Apoio à Agricultura Digital Brasileira (2023). In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO (SBSI), 19. , 2023, Maceió/AL. Anais [...]. Porto Alegre: SBC, 2023. p. 90-92. DOI: https://doi.org/10.5753/sbsi_estendido.2023.229255.

Palavras-chave: opensoils; inteligência artificial; sustentabilidade.

