Combining Hard and Soft Voting Machine Learning Algorithms for Soil Classification.

*Sally Ndikum Joseph Pemberton

Olushina Olawale Awe
African Institute for Mathematical Sciences, (AIMS)
Limbe, Cameroon
*sally.ndikum@aims-cameroon.org

Abstract

Civil engineers need to possess knowledge about soil properties and structure through soil classification to construct reliable and long-lasting structures. However, traditional soil classification methods are both expensive and time-consuming. Recently, machine learning has become increasingly popular in solving complex problems in Geotechnical Engineering, leading to novel approaches for automating soil classification. This research evaluates the effectiveness of various machine learning algorithms, including Multinomial Logistic Regression (MLR), Gaussian Naive Bayes (GNB), Extreme Gradient Boosting (XGBoost), Random Forest, and Artificial Neural Network-Multilayer Layer Perceptron (ANN-MLP), in classifying soils. The study also implemented Hard and Soft Voting Ensemble Learners. Each model was quantitatively evaluated and compared using various metrics. Empirical findings suggest that all models are effective in classifying soils, with the hard voting model outperforming the others.

1 Introduction

Feasibility and preliminary studies are conducted before any construction project to assess its viability. Soil classification is a crucial aspect of feasibility studies to determine whether the soil is suitable for the construction project. The main objective of soil classification is to group soils with similar characteristics and properties into relevant categories and subcategories based on their identified properties. This provides the civil engineer with valuable insights into the type and structure of the soil [3]. Soil classification is extremely important from an engineering standpoint [2], as a lack of information about the soil type and structure can lead to structural failures, such as building collapses and road degradation. Moreover, this study aligns with SDG9, which focuses on industry, innovation, and infrastructure. The goal aims to enhance research and upgrade industrial technologies for resilient building and infrastructure design and management. The use of machine learning in civil engineering, as applied in this study, supports this objective. Thus, Geotechnical engineers are playing increasingly important roles in the civil engineering sector and are using machine learning techniques in various domains to address complex issues in the field.

The process of soil classification is typically done by analyzing soil properties such as particle size and plasticity, which requires specialized knowledge and laboratory procedures. However, this makes it a costly and time-consuming process that requires a team of skilled geotechnical engineers. Traditional methods of soil classification, which involve cone penetration tests, trial pit testing, and laboratory tests, rely solely on human-based approaches and are not able to capture intricate patterns in soil data. To address the challenges posed by traditional methods in soil classification and improve

5th Deep Learning Indaba Conference (DLI 2023).

their effectiveness, this study aims to explore and compare various machine learning algorithms for automating the process. As such, this study is essential.

The study aims to develop and evaluate the performance of various machine learning models using a specific set of soil data.. The models employed consist of Gaussian Naive Bayes (GNB), Multinomial Logistic Regression, ANN (Multilayer Perceptron), Random Forests (RF), and Extreme Gradient Boosting (XGBoost), and a hybrid approach that employs ensemble techniques of hard and soft voting. The selection of these algorithms offers a diverse set of models that can handle different aspects of the soil classification task. The contribution lies in utilizing a new dataset comprising Cameroonian soils and employing a variety of ML techniques to classify them.

Recent studies have demonstrated the gradual development and application of various machine learning techniques to improve performance and accuracy in soil classification. For instance, Nguyen et al.[6] employed Support Vector Classification (SVC), Multilayer Perceptron Artificial Neural Networks (ANN), and Random Forest models to classify 4888 soil samples into five classes using 15 selected input features. Their analysis revealed that all three models performed well, with the SVC model being the most effective. Also, Pham et al.[8] proposed new models for soil classification based on ANN, Decision Tree, and AdaBoost algorithms. They used AdaBoost models to classify 440 soil samples, with their models based on two enhanced tree algorithms and a neural network. Their findings indicated that the Adaboost model was successful in accurately classifying the soils, even though tree-based algorithms are not frequently used in this area.

2 Methodology

The soil classification dataset was created by collecting 242 samples from a road construction project in Bamenda, Cameroon, and testing them in a well-equipped geotechnical laboratory with soil identification tests, including sieve analysis and Atterberg's limits. The resulting identifiable soil parameters were used as input variables for machine learning models. The dataset has 17 variables and 6 classes representing different soil types. Python libraries, such as Scikit-learn and Numpy, were used for expl 0/*6-y analysis, model training, and evaluation.

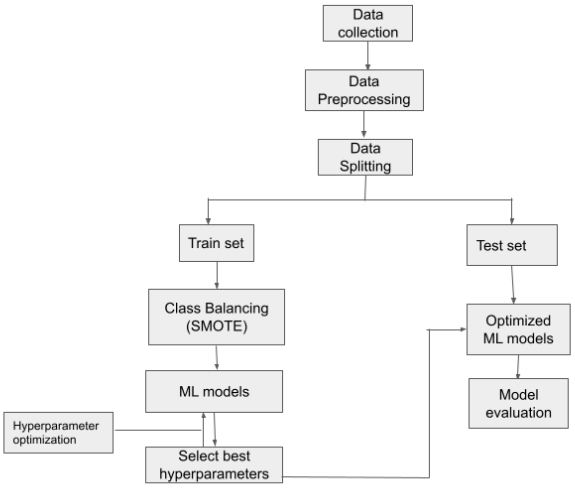

Figure 1: Proposed Methodology

2.1 Data Preprocessing

2.1.1 Feature Selection

RFE was used to select important features for prediction. Feature importance scores were determined using the random forest algorithm, and the five most important features were selected. These features

are commonly used in soil classification, including retaining 2mm and 0.4mm sieves, liquid limit (LL), plastic limit (PL), and plasticity index (PI).

2.1.2 Class Imbalance

To address the imbalance in soil classes observed during sampling where some classes were more frequent while others were rare due to the natural variation in soil formation within a meter, SMOTE (Synthetic Minority Oversampling Technique) was used. This technique generates synthetic samples between the nearest neighbors of observations in the minority class, thereby increasing samples in the minority classes and producing an equal distribution of classes in the dataset.

2.2 Classification with Machine learining Algorithms

The dataset used in this study has a multi-class categorical target variable with a label (soil class). Thus, classification algorithms had to be implemented to formulate a ML model for the prediction of the target variable (soil class). The algorithms used in this study include:

Random Forests: A tree-based technique which is a collection of many decision trees which are built in parallel using the bagging algorithm. Bagging begins by randomly selecting subsets of data from the original dataset with replacement. These are referred to as bootstrapped datasets. On each of these bootstrapped datasets, a defined number of decision trees are independently trained with subsets of the selected features.

Extreme Gradient Boosting An enhanced gradient boosting algorithm consisting of an ensemble of weak learners(e.g decision trees) built sequentially, such that the errors of the previous learner is used to correct the errors of the subsequent learner, thus improving the final model's accuracy.

Gaussian Naive Bayes: GNB algorithm is based on the Bayes Theorem and assumes conditional independence between features and assumes that the features are continuous and follow a Gaussian distribution.

Multinomial Logistic Regression: MLR finds the probability of a sample to belong to each possible class using the softmax function. The softmax function maps a vector of real values to a vector of probabilities that sum to 1. This produces a probability distribution over the classes, where each probability represents the likelihood that the input belongs to that class.

ANN-MLP: Artificial neural networks function similarly to the human brain[1]. Perceptrons are the basic unit of an ANN, consisting of an activation function, inputs, and outputs, organized in layers to form a Multilayer Perceptron (MLP). ANNs can identify complex and non-linear relationships between input and output neurons, and are used in machine learning[7].

Hard and Soft voting: To improve predictive performance, machine learning models are combined using the ensemble techniques of Hard and Soft voting [4]. In Hard voting, the final prediction is based on the majority vote from each model, while in Soft voting, the prediction with the highest average probability is selected by averaging the probabilities of each prediction from each model [5].

During model training, Hyperparameter optimization of each model was done using Gridsearch Cross-validation. This technique is used to determine the best combination of hyperparameters for a machine learning model by creating a grid of hyperparameter values and training a model for each grid combination of hyperparameters. The dataset was split in the ratio 70/30.

2.3 Performance metrics of the classifiers

In order to understand which model/algorithm performs best for a given dataset, performance metrics are required. In this study, some performance metrics used to evaluate the performance of the models. They include Balanced Accuracy, precision, Recall, F1-score, Matthews Correlation coeffficient.

3 Results

3.1 Impact of data balancing

When the model accuracies are evaluated before and after data balancing with SMOTE, it is clear that dealing with class imbalance improves model performance. SMOTE increases the model's accuracy by providing more training data for the minority class, which helps the machine better understand how to distinguish between the classes. The accuracy changes of each model before and after data balancing are shown in Table 1.

Table 1: Effect of Balanced data on accuracy of Soil Classification Algorithms using SMOTE.

Model	Bal. Accuracy before SMOTE	Bal.Accuracy after SMOTE	% Increase
Random Forest	86.9	87.6	+0.7%
XGBoost	90.9	91.9	+1%
Naive Bayes	74.7	77.6	+2.9%
MLR	91.7	93.2	+1.5%
MLP	82.7	91.9	+9.2%

3.2 Comparison of model performance

When comparing the performance of different models, we use 5 different random seeds and obtain the average of the results from each random seed. This is to maintain fairness and obtain reproducible results. This is because the random initialization of weights or the random shuffling of data can alter the performance of some models. These models are evaluated on the test-set consisting of 72 samples, which make up 30% of the whole dataset. The results for the models are displayed in Table 2, including their individual performances on the test set.

Table 2: Comparative Evaluation of ML models

Model	Balanced Accuracy	Precision	Recall	F1-score	MCC
Random Forest	87.6	89.7	86.7	88.0	80.6
XGBoost	91.9	95.3	91.7	93.0	89.2
Naive Bayes	77.6	73.7	76.3	73.3	59.9
MLR	91.7	90.7	92.3	91.0	**89.8**
MLP	90.3	93.3	91.7	92.3	86.3
Hard voting	**92.8**	**94.4**	**93.1**	**93.7**	87.3
Soft voting	92.1	92.3	91.0	91.3	88.2

4 Conclusion

In Civil Engineering, knowing soil classes is crucial for constructing reliable structures, and machine learning provides novel approaches for improving the efficiency of soil classification. The machine learning models exhibited high balanced accuracies ranging from 87% to 93%, except for the Naive Bayes model, which had a lower accuracy across all classes. The Random Forest and XGBoost models performed well, with high precision, recall, and F1-scores. The MLR model had the highest MCC score and a balanced accuracy of 91.7%. The MLP model, a type of ANN, achieved a balanced accuracy of 90.3%. Finally, the Hard voting algorithm provided the highest balanced accuracy of 92.8%, likely due to its combination of multiple models to reduce losses.

Voting ensemble algorithms can be integrated into software programs for soil classification. The application of machine learning in Geotechnical Engineering is a continuously growing field, thus it is recommended to explore different machine learning algorithms including ensembling methods and ANN architectures for soil classification.

Acknowledgments and Disclosure of Funding

The authors sincerely thank the company; Bambuiy Engineering Services and Techniques (B.E.S.T) for providing the raw data used in this study. No funding was received for this paper.

References

[1] K. J. Atmaja, I. B. N. Pascima, I. M. D. P. Asana, and I. G. I. Sudipa. Implementation of artificial neural network on sales forecasting application. *Journal of Intelligent Decision Support System (IDSS)*, 5(4):124–131, 2022.

[2] Y. Aydın, Ü. Işıkdağ, G. Bekdaş, S. M. Nigdeli, and Z. W. Geem. Use of machine learning techniques in soil classification. *Sustainability*, 15(3):2374, 2023.

[3] R. T. Chandan and R. Thakur. An intelligent model for indian soil classification using various machine learning techniques. *International Journal of Computational Engineering Research (IJCER)*, 33(2250):3005, 2018.

[4] S. Kumari, D. Kumar, and M. Mittal. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, 2:40–46, 2021.

[5] A. Manconi, G. Armano, M. Gnocchi, and L. Milanesi. A soft-voting ensemble classifier for detecting patients affected by covid-19. *Applied Sciences*, 12(15):7554, 2022.

[6] M. D. Nguyen, R. Costache, A. H. Sy, H. Ahmadzadeh, H. Van Le, I. Prakash, and B. T. Pham. Novel approach for soil classification using machine learning methods. *Bulletin of Engineering Geology and the Environment*, 81(11):468, 2022.

[7] Q. H. Nguyen, H.-B. Ly, L. S. Ho, N. Al-Ansari, H. V. Le, V. Q. Tran, I. Prakash, and B. T. Pham. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021:1–15, 2021.

[8] B. T. Pham, M. D. Nguyen, T. Nguyen-Thoi, L. S. Ho, M. Koopialipoor, N. K. Quoc, D. J. Armaghani, and H. Van Le. A novel approach for classification of soils based on laboratory tests using adaboost, tree and ann modeling. *Transportation Geotechnics*, 27:100508, 2021.

A Appendix

The Python code used for this study is accessible with this link.

`https://drive.google.com/drive/folders/1oQ3lGF6OabyXm5oMm5GEzumY8ELBGO4y?usp=share_link`