Soil Fertility Prediction

Soil fertility prediction involves using data and algorithms to estimate the nutrient content and health of soil for optimized agricultural practices.

Author

AISHWARYA. S
1AT22CD003
CSE (Data Science)

INTRODUCTION

Soil fertility is essential for sustainable agriculture and food production. This project aims to predict soil fertility using machine learning techniques, analyzing key parameters such as nitrogen (N), phosphorus (P), potassium (K), pH levels, and micro-nutrients like zinc (Zn) and copper (Cu). By classifying soil as "Fertile" or "Not Fertile," the project provides actionable insights for better soil management.
The predictive model leverages data-driven analysis to assess soil health and suggest improvements to enhance crop yields. This initiative supports precision agriculture, helping farmers optimize resources and address the growing global demand for food security efficiently.

DATA COLLECTION

The soil fertility dataset comprises 881 samples and 13 features, encompassing various soil properties like nutrient content (N, P, K), soil acidity (pH), electrical conductivity (EC), and presence of essential elements like sulphur (S), zinc (Zn), iron (Fe), copper (Cu), manganese (Mn), and boron (B). Soil samples are categorized into two fertility classes: "Not Fertile" (1) and "Fertile" (0).

DATA PREPARATION

Data Check: The dataset was checked for null and duplicate values, ensuring data integrity.
Encoding: Since the target variable has two classes (Fertile and Not Fertile), label encoding was applied, assigning numerical values to each class that is Fertile as "0" and Not fertile as "1".
Scaling: Min-max scaling was performed on numeric features to bring them within a fixed range (0-1), aiding model convergence and performance.
Outlier Detection and Removal: Boxplot analysis revealed outliers in the nitrogen (N) and phosphorus (P) features. Outliers were removed to improve the robustness of the dataset.

METHODOLOGY

The models considered include K-Nearest Neighbors (KNN), Decision Trees (DT) with and without bagging, Random Forest, Logistic Regression, and Bagging Classifier.
Model Training and Evaluation:

- The dataset was split into training and testing sets using a 70-30 ratio.
- Model performance metrics such as confusion matrix, classification report, and accuracy score were computed for each model.
- Grid search was employed to optimize hyperparameters for Decision Trees, ensuring enhanced model performance

IMPLEMENTATION

The project began with data preprocessing to handle missing values and normalize features. Soil properties like nitrogen, phosphorus, potassium, pH, and micro-nutrients were analyzed for soil fertility prediction. The dataset was split into training and testing sets for model evaluation. Multiple algorithms, including Decision Trees, Random Forest, and Logistic Regression, were used, with performance assessed through accuracy, precision, recall, and F1-score. Hyperparameter tuning optimized the final model, which classifies soil as "Fertile" or "Not Fertile." The system offers actionable insights for farmers, improving soil management and agricultural productivity.

RESULT

K-Nearest Neighbors (KNN) achieved an accuracy of 89%. Decision Tree (DT) achieved an accuracy of 92%. Random Forest attained an accuracy of 94%. Logistic Regression reached an accuracy of 91%. Bagging was performed to assess potential overfitting for the Decision Tree model, and the accuracy for both the original and bagged models remained almost the same. Considering the overall performance, Random Forest exhibited the highest accuracy among the tested algorithms.

CONCLUSION

Random Forest emerged as the top performer with an accuracy of 94%. While Decision Tree, Logistic Regression, and KNN also demonstrated respectable accuracies, Random Forest's robust performance suggests its suitability for soil fertility prediction tasks. Bagging, although applied specifically to the Decision Tree model in this study, offers a valuable technique for assessing overfitting and enhancing generalization across unseen data.