**REVIEW COMMENTS**

**Manuscript Title: Soil Analyzer - Analyse the soil fertility with the Essential Nutrients Present in the Soil**

**Decision: Accepted with Major Revision**

**Review Comments 1:**

1. **Soil analyzer -analyzing the soil fertility with the essential nutrients present in the soil is the proposed title of this paper.**

The title has been updated to "Machine Learning-Based Soil Fertility Analysis for Informed Decision-Making".

**2. How to achieve the data cleaning process?**

Data cleaning process I implemented:

1. **Handling Missing Values:**

* Identified and replaced zero values with NaN in relevant columns.
* Used mean and median imputation to fill missing values in specific columns, considering the distribution of the data.
* Applied imputation separately for each feature.

1. **Standardization:**

* Removed unnecessary columns ('id', 'label', 'SFI') from the features (X) and separated the target variable (y).
* Used StandardScaler from sklearn to standardize the feature values.

These steps ensure that your dataset is prepared for machine learning modeling, with missing values addressed and features standardized for consistent scale.

**3. How to balance the dataset?**

Since the dataset obtained from Kaggle, prepared by G. B. Pant University of Agriculture and Technology, was already balanced, no additional balancing processes were required.

**4. How to achieve the normalization process?**

Normalization was accomplished to ensure that the various features in the dataset were on a consistent scale. This is crucial for machine learning models as it prevents features with larger numerical ranges from dominating the learning process. The normalization process involved using the StandardScaler from the scikit-learn library. The selected features, excluding unnecessary columns such as 'id,' 'label,' and the target variable 'SFI,' were scaled to have a mean of 0 and a standard deviation of 1.

**5. Dataset details are inadequate.**

The dataset, meticulously prepared by G. B. Pant University of Agriculture and Technology, encompasses a diverse array of soil parameters, including pH, electrical conductivity, organic carbon, and micronutrient concentrations. This extensive dataset, containing a total of 2738 instances, significantly contributes to understanding the dataset's size. It provides a robust foundation for our analysis, allowing for comprehensive insights into soil characteristics and fertility.

**6. How to predict the soil fertility?**

To predict soil fertility, we implemented the Random Forest machine learning model, an ensemble learning algorithm that analyses various soil parameters to predict the Soil Fertility Index (SFI). The model, through multiple decision trees, considers features like pH, nutrient levels, soil texture, organic carbon, and electrical conductivity. Its strength lies in handling complex relationships, capturing feature importance, and providing reliable predictions. With the predicted SFI score, farmers receive a clear assessment on a scale from 1 to 10, guiding them in optimizing fertilizer application, minimizing costs, and promoting soil health. This precise approach contributes to sustainable agriculture, preventing the overuse of fertilizers that could potentially harm crops and the environment.

**7. How to improve the performance?**

In refining the soil fertility prediction model, strategic measures were employed, encompassing feature engineering, hyperparameter tuning, and rigorous model evaluation. Feature engineering involved a meticulous selection of relevant soil parameters, creating composite features to enhance the model's understanding of intricate data interactions. Concurrently, hyperparameter tuning fine-tuned the Random Forest algorithm for optimal predictive performance. Robust model evaluation was ensured through meticulous cross-validation techniques, testing adaptability across diverse soil conditions. To sustain relevance, the model underwent continuous scrutiny, with ongoing retraining and integration of new data to capture evolving patterns. This resulted in an optimized soil fertility prediction model, demonstrating heightened performance in real-world agricultural scenarios.

To maintain efficacy, a proactive data maintenance strategy was implemented, recognizing the dynamic nature of soil nutrient requirements influenced by evolving agricultural practices. Regular infusion of updated soil nutrient data ensures model relevance, aligning with agriculture's dynamic nature, and sustaining accuracy as a valuable tool for farmers in informed decision-making about fertilizer application and sustainable soil management.

**8. How to maintain dataset integrity?**

Maintaining dataset integrity in the open-source web application involves a careful balance between accessibility and privacy. As the webpage is freely accessible to any farmer, data collection from users is deliberately avoided to uphold privacy standards. This approach ensures the integrity of the dataset by preventing any interference or alteration of user data, emphasizing transparency, and respecting user privacy rights within the open-source framework.

**9. How to perform the data pre-processing process?**

In the data pre-processing phase, the commitment to data integrity guides the meticulous preparation of the dataset for meaningful soil analysis. Initially, gaps and missing values are addressed by identifying and replacing zero values with nan in relevant columns. To fill these gaps, mean and median imputation techniques are applied separately to each feature, considering the data distribution. Simultaneously, potential outliers are identified and treated using appropriate methods to safeguard the integrity of the soil analysis project.

Furthermore, the data pre-processing journey involves the crucial step of standardizing features. Unnecessary columns ('id', 'label', 'SFI') are removed, and the target variable (y) is separated from the features (X). The scikit-learn library's standardscaler is then employed to standardize feature values, ensuring a consistent scale across various parameters. This comprehensive approach, encompassing missing value handling and standardization, establishes a robust foundation for subsequent machine learning modeling. The prepared dataset, sourced meticulously from a reputable institution, is now well-equipped for the Soil Analyzer project's data analysis, laying the groundwork for precise and meaningful soil fertility assessments.

**10. The title should include a technical term. Avoid using personal pronouns throughout the article content.**

The title has been updated to " Machine Learning-Based Soil Fertility Analysis for Informed Decision-Making".

**Review Comments 2:**

**1. Soil Analyzer - Analyse the soil fertility with the Essential Nutrients Present in the Soil is the presented research work title.**

The title has been updated to " Machine Learning-Based Soil Fertility Analysis for Informed Decision-Making".

**2. Authors are suggested to cite relevant references for the discussion given in the introduction section.**

Citations have been appropriately integrated to substantiate the information presented in the introduction section, ensuring the inclusion of relevant references to support the context and background of the study.

**3. The limitations observed in the existing works should be summarized in the last part of the related works section.**

The dynamic nature of soil micronutrient requirements and the need for further analysis by agricultural experts to determine the precise amount of fertilizer represent notable limitations in the project. The experimental scores generated by the model provide a valuable initial assessment, but the evolving nature of soil conditions underscores the importance of expert validation for accurate and context-specific recommendations. This limitation highlights the collaborative nature of implementing machine learning models in agriculture, where domain expertise remains crucial for refining and contextualizing the model's predictions.

**4. How the dataset is balanced? The technique used for data balancing should be discussed.**

The dataset obtained from Kaggle, prepared by G. B. Pant University of Agriculture and Technology, was inherently balanced, eliminating the need for additional balancing techniques. The original data collection process ensured a proportional representation of instances across different classes, maintaining balance without requiring specific balancing procedures. This balanced distribution is advantageous for machine learning model training, fostering effective pattern learning across all categories.

**5. Why random forest has been selected for the proposed work? Algorithm selection should be validated.**

The selection of the Random Forest algorithm for the proposed work was validated through a comprehensive evaluation based on key performance metrics, including accuracy, precision, recall, F1 score, and mean absolute error. The comparison involved other prominent algorithms, namely Support Vector Machine (SVM) and Decision Tree.

In assessing classification performance, Random Forest exhibited superior accuracy, precision, recall, and F1 score compared to SVM and Decision Tree. This outcome underscores the algorithm's effectiveness in achieving a well-rounded and balanced classification across different soil fertility categories. Specifically, Random Forest demonstrated an impressive accuracy score of 99.66%, emphasizing its precision in categorizing soil fertility. Moreover, for regression tasks, the Random Forest algorithm demonstrated its superiority by minimizing the mean absolute error, showcasing a remarkable value of only 0.03 degrees. The algorithm's ensemble learning approach, which combines multiple decision trees, enables it to capture complex relationships within the dataset, resulting in more accurate predictions of continuous variables such as the Soil Fertility Index (SFI) in the context of this project.

The Random Forest's robustness, ability to handle intricate data patterns, and consistent performance across classification and regression tasks make it the optimal choice for the proposed work. These findings highlight the algorithm's suitability for soil fertility prediction and underscore its potential to provide reliable insights for informed agricultural decision-making.

**6. Essential formulations for the machine learning model should be included with a brief discussion.**

The essential formulations for the machine learning model in the research paper have been systematically detailed in the methodology section. The step-by-step discussion outlines the key formulations involved in predicting soil fertility using the Random Forest algorithm. It encompasses the formulation of the Soil Fertility Index (SFI) with specific attention to the constraints applied to each feature, referencing global data. The comprehensive methodology section provides a clear and detailed understanding of how the machine learning model processes soil parameters, such as pH, nutrient levels, soil texture, organic carbon, and electrical conductivity, to predict the SFI. This inclusive discussion ensures transparency and clarity in the formulation of the model, contributing to the reproducibility and understanding of the research.

**7. Authors are suggested to include a detailed performance metrics analysis of the proposed model.**

The research paper incorporates a thorough performance metrics analysis of the proposed model. Utilizing accuracy, precision, recall, F1 score, and mean absolute error, the evaluation provides a comprehensive assessment of the model's predictive capabilities. The metrics serve to validate the effectiveness and reliability of the Random Forest algorithm in predicting soil fertility. This detailed analysis enhances the credibility of the research findings and underscores the robust performance of the proposed model across various dimensions, contributing to the overall rigor of the study.

**8. It is suggested to include a comparative analysis of the proposed model and existing applications and discuss the observations in detail**

The research paper features a comparative analysis between the proposed model and existing applications. Through this analysis, key observations and distinctions are meticulously discussed. By evaluating the strengths and weaknesses of the proposed model in comparison to existing applications, the paper provides valuable insights into the novelty and efficacy of the developed Soil Analyzer. This comparative discussion contributes to the overall depth of the research, offering a nuanced understanding of how the proposed model stands out in the landscape of soil fertility prediction applications.