I. PROPOSED METHODOLOGY AND IMPLEMENTATION

The proposed method uses import models like Linear Regression, Support Vector Regression (RBF), Support Vector Regression (Poly), Support Vector Regression (Linear), Decision Tree Regression, Random Forest Regression, Gradient Boosting Regressor and Stacking Regressor for data collected, pre-processed, visualization etc. The primary goal is to analyze and predict soil properties as well as conditions based on collected data. This can include parameters such as moisture content, nutrient levels, pH, and other relevant factors. The model is designed so that it helps farmers optimize resource use, improve crop yields, and reduce environmental impact through precise decisions on irrigation, fertilization, and crop selection based on accurate soil data. We didn't use Linear Regression or SVM due to the fact that the result we got was not up to the mark. The R2 score is very low as well as MSE (Mean Squared Error) was high too. The implementation phase marks the transition from theoretical groundwork to practical application, bringing the devised methodology to life with a focus on the real-world integration of Machine Learning (ML) in soil monitoring. For Selecting the best model we'll be following approach shown in Fig.1.

A. *Data Collection:* This dataset contains the various elements found in the soil, for instance, organic matter, various nitrogen compounds, potassium, sodium, sulphates, boron, etc., It also contains various soil properties like pH. The target of this data is set to predict the vegetation cover which is the percent vegetative cover. The higher the vegetation cover, the higher is the fertility of the soil for crops. Vegetation cover is calculated in percentage from 1 to 100, so, it becomes a regression task. To achieve the results various regression methods are applied and performance of each model is analyzed [17].

B. *Data Preprocessing:* We imported the data into our jupyter workbook, then viewed the first few rows of the dataset to see the type of data we are working with. Then we checked and removed unnecessary columns and rows that do not contain data and rename column names and units of each attribute. We create a new dataset from this and save it as soil_data. Then we check for missing values and fill the rows with 0.Using a simple imputer from sklearn, we transformed the dataset and changed 0 values to the median of other row's values. We check again the number of NaN values in each column to see if there are still any missing values. Then saved the processed data into a csv file as processed_data_set.

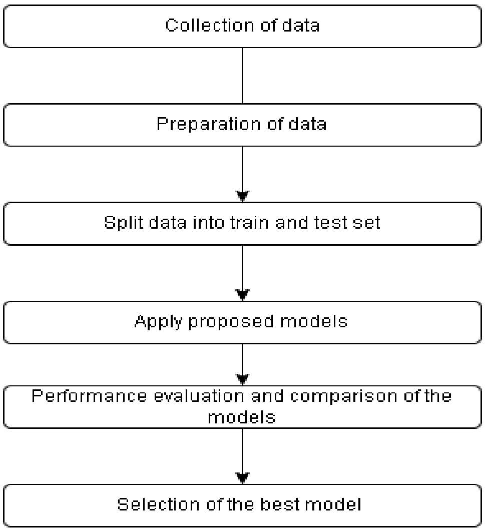

Collection of data
Preparation of data
Split data into train and test set
Apply proposed models
Performance evaluation and comparison of the models
Selection of the best model

Fig. 1. Flow Chart of Propsoed Methodology

C. *Implementing Machine Learning Models:* We have imported models like Linear Regression, Support Vector Regression (RBF), Support Vector Regression (Poly), Support Vector Regression (Linear), Decision Tree Regression, Random Forest Regression, Gradient Boosting Regressor and Stacking Regressor.

D. *Performance Evaluation ,Comparing and Selecting Best Models*: We'll import models and compare them using R2 scores and MSE(mean square error) and selecting on the basis of that.

II. RESULT ANALYSIS

A. Data Visualisation For Vegetation Cover

Fig 2.1,2.2,2.3 are the pair plots which explain the relationship between vegetation covers and all the other factors. Pair plots are used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our data-set.

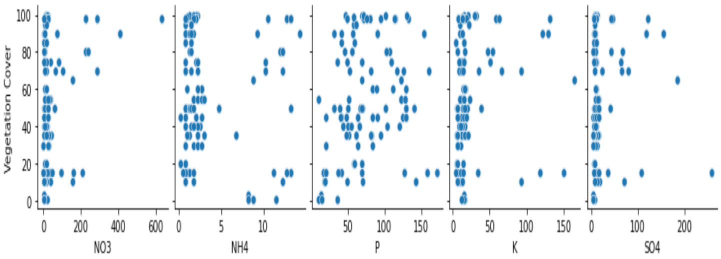


Fig.2.1. Pair Plot of vegetation cover with other factors

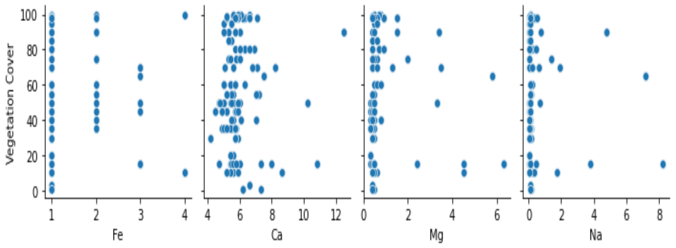


Fig.2.2. Pair Plot of vegetation cover with other factors

.

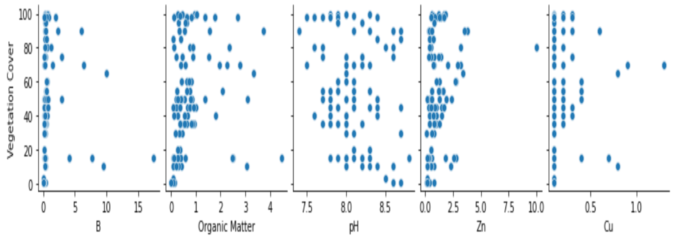
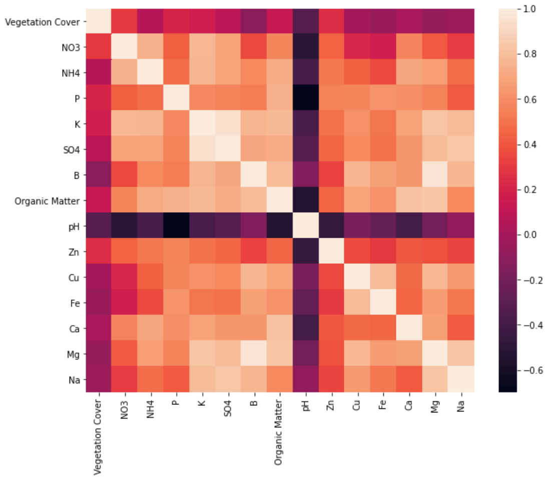


Fig.2.3. Pair Plot of vegetation cover with other factors

.

B. Data Visualisation (Heatmap)

Using seaborn, we’ll make a heat map using all the columns as shown in the map below. Fig.3 is a heatmap where we can understand how the fertility of the soil depends on all the factors and variables from the dataset. A heat map consists of values exhibiting different shades of one colour for individual values to be plotted. It is a data visualization technique that displays the magnitude of a phenomenon as colour in two dimensions. From the map we can conclude that vegetation cover is dependent on pH the most. That means the pH is responsible the soil to be fertile or not the most.



C. *Model Comparison*

We have compared the models according to their R2 scores and MSE (mean square error). As we can see that all the models give us an exceptional score as well as MSE. To choose the best model, we’ll have to take a look over the R2 scores and their MSE as well.

- The Stacking models, especially when using Gradient Boosting Regressor as the meta-estimator, have the highest R2 scores, indicating better overall performance.
- Gradient Boosting Regressor alone also performs very well, with high R2 scores and the lowest MSE among all the models.
- Random Forest Regressor follows closely in performance with slightly lower R2 scores than Gradient Boosting Regressor.
- Decision Tree Regressor has the lowest performance among the models, with significantly lower R2 scores and the highest MSE.

We are only using the Stacking model with Gradient Boosting Regressor as the meta-estimator as it has a better performance than the Stacking model with Random Forest Regressor as the meta-estimator. Table II Compares the models on the basis of R2 score and Mean Square Error, showing how each model works and predicts.

TABLE II
COMPARISON BETWEEN MACHINE LEARNING MODELS

S. No	**MODEL**	**R2_SCORE (TRAIN DATA)**	**R2_SCORE (TEST DATA)**	**MSE**
1	Decision Tree Regressor	0.893	0.704	6.709
2	Random Forest Regressor	0.963	0.927	6.111
3	Gradient Boosting Regressor	0.969	0.939	5.859
4	Stacking(meta=RFR)	0.99912	0.971	6.856
5	Stacking(meta=GBR)	0.998	0.972	6.773

D. *Data Visualisation Of R2 Scores And MSE*

Fig 4.1,4.2 and 4.3 plotted to demonstrate and compare the scores and MSE for all models. In figure 4.1, after comparing we can see that Stacking model with meta regressor as GBR model provides the highest R2 score, meaning that the proportion of the variance in the dependent variable (fertility) can be explained by the independent variables (features) in the model. The second best behind it comes the GBR model with slightly less R2 score but still an amazing result.

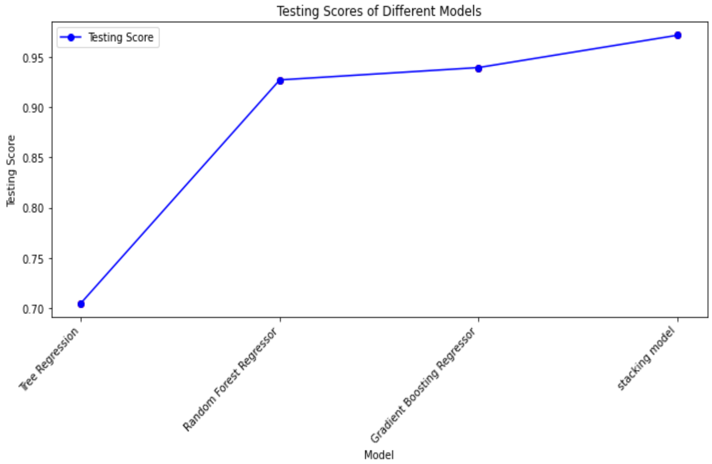


Fig.4.1. Comparing Training R2 Scores

Similarly in Fig. 4.2 with the testing dataset, we can see that Stacking model with meta regressor as GBR model still provides the highest R2 score, with GBR model coming second with slightly less R2 score.

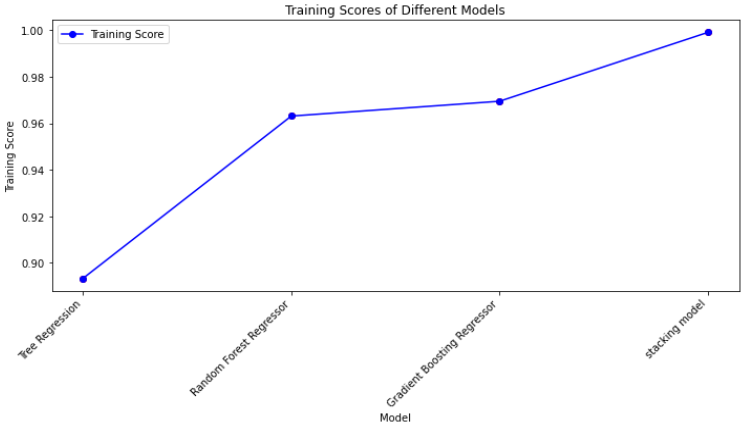


Fig.4.2. Comparing Testing R2 Scores

MSE in Fig. 4.3 measures the average squared difference between the predicted values and the actual values. Lower the MSE values, better the model is at predicting the target variable, as it means the predictions are closer to the actual values. Here, the best result is shown by our GBR model with the least MSE score followed by the Random Forest model.

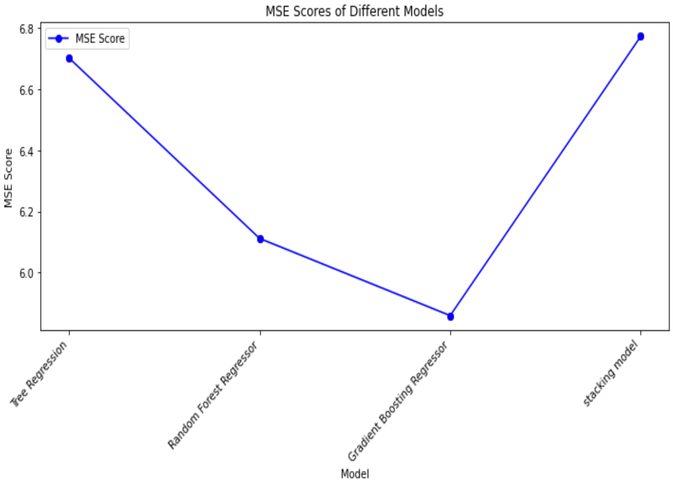


Fig.4.3. Comparing Testing MSE Scores

E. Data Visualisation (True Vs Predicted and Residual Vs Predicted)-

In Fig. 5, we have plotted a graph between True values and predicted values for all the models. It helps us to visualize how each model works and how each of them is accurate. The closer the points are in a straight line better the accuracy is of the model. As we can observe, the stacking model and the gradient boosting points are closer to the linear line that we have drawn in the map. Hence, we can say that the two models mentioned previously are more accurate than the other models.

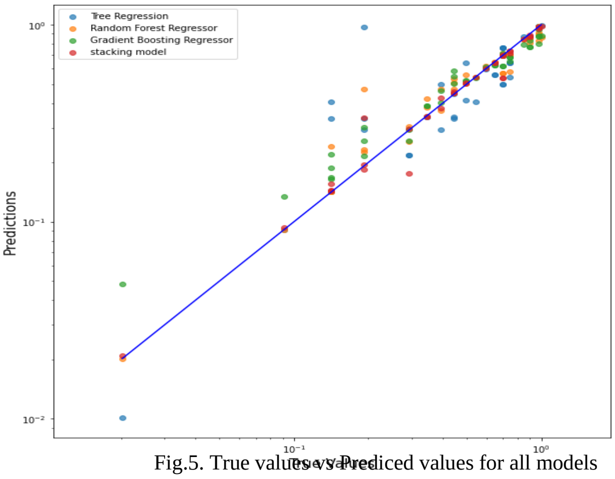


Fig.5. True values vs Prediced values for all models

In Fig.6, we have plotted the graph for residual values and predicted values. For all the models the residuals appear randomly scattered around the horizontal line at y=0, it suggests that the models are performing well and capturing the variation in the data. We can also observe that the scattering in stacking and GBR is much less than other models, hence they seem to be performing much better than other models.

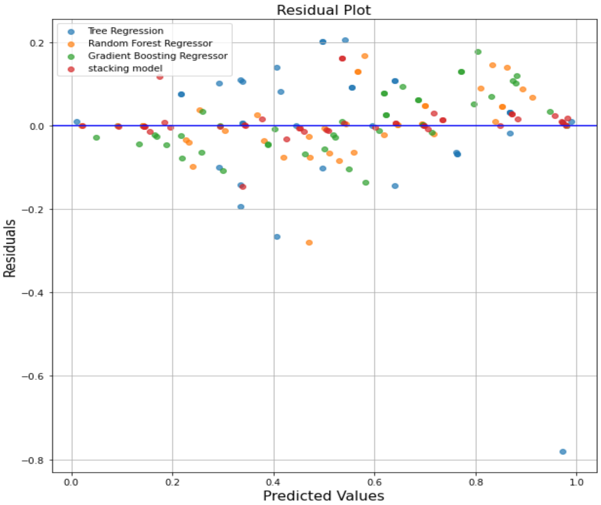


Fig.6. Residual plot between reidual and predicted values

F. K-FOLD (5-fold) VALIDATION

In 5-fold cross-validation, the dataset is divided into five equal-sized folds. The learning algorithm is then trained and evaluated five times. During each iteration, one-fold serves as the validation set while the remaining four folds are used for training. This process allows for comprehensive evaluation, ensuring that every data point gets a chance to be in the validation set at least once. By averaging the performance metrics, a more robust estimate of the model's performance is obtained, which helps us in assessing how well a model generalizes to unseen data. This method is particularly beneficial when the dataset is limited in size, as we have, it maximizes the use of available data for both training and evaluation, reducing the impact of variability in a single train-test split. In Fig.7, We have plotted the graph of each fold for all the models where y-axis is predicted values and x-axis is the true values. We also calculated the mean MSE and the std MSE (standard deviation MSE). Here, we can conclude that the stacking model seems to give the best accuracy and MSE as predicted and the true values are closer in a linear *line.*

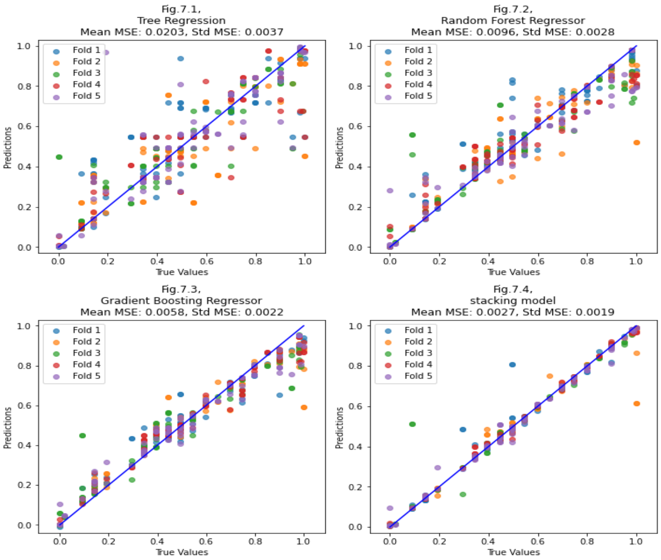


Fig.7.1,
Tree Regression
Mean MSE: 0.0203, Std MSE: 0.0037

Fig.7.2,
Random Forest Regressor
Mean MSE: 0.0096, Std MSE: 0.0028

Fig.7.3,
Gradient Boosting Regressor
Mean MSE: 0.0058, Std MSE: 0.0022

Fig.7.4,
stacking model
Mean MSE: 0.0027, Std MSE: 0.0019

Fig.7. Comparison of all the folds(5), K-fold valifation for all models

In Table III, we have compared the mean MSE and the Std MSE between the models after K-fold (5) validation. We can conclude that stacking model is the best performing as it has the least mean MSE and std MSE. This means this model would give us the most accurate predictions for the fertility of the soil.

TABLE IIII

MEAN MSE, STANDARD DEVIATION MSE COMPARISON FOR ALL MODELS

S.NO.	Model	Mean MSE	Std MSE
1	Tree Regression	0.0203	0.0037
2	Random Forest	0.0096	0.0028
3	Gradient Boosting	0.0058	0.0022
4	Stacking	0.0027	0.0019

In Table IV, we have made a comparison between previous work and our work to demonstrate how our study provides a better model with better accuracy.

TABLE IV

COMPARISON WITH EXISTING WORK

S. No.	Technique Used	Accuracy (%)	Source
1	Random Forest Regression	92.7	[15]
2	Decision Tree	8	[16]
3	Stacking model meta (Gradient Boosting Regressor)	99	Proposed Work

After comparing and analyzing all the models, we analyzed that the best suitable algorithm to use would be GBR (Gradient Boosting Regressor) as it performs excellently in all the areas. The stacking model with meta as GBR comes as the close second best with the highest R2 score but similar MSE when compared to other models.