Soil NPK Prediction Using Multiple Linear Regression

[1]Madhumathi R,
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore, India
madhumathi.r@srec.ac.in

[2]Arumuganathan T,
ICAR-Sugarcane Breeding Institute,
Coimbatore, India
arumuganathan@gmail.com

[3]Shruthi R,
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College,
Coimbatore, India
shruthi.1801217@srec.ac.in

[4]Raghavendar S,
Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College, Coimbatore, India
raghavendar.2001154@srec . ac.in

2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) | 978-1-6654-0816-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACCS54159.2022.9785338

***Abstract*—Soil nutrients are the important parameter which contributes a major role in healthy plant growth. In soil, the presence of three macro nutrients namely Nitrogen (N), Phosphorus (P), and Potassium (K) are essential for proper crop growth. Without having adequate knowledge about nutrient levels present in soil, farmers apply large quantity of fertilizers in their field. This leads to depletion or enhancement of nutrient content in the soil and it degrades the soil fertility. Laboratory soil test is time consuming and it involves addition of many chemical reagents. Hence in this paper, the soil macronutrients are predicted using Multiple Linear Regression (MLR) technique. It is a statistical method where several independent or explanatory variables are used to predict the dependent or response variable. MLR technique is formulated to determine a mathematical relationship among several parameters. It shows the relationship between dependent and independent variables. In this technique, soil parameters like nitrogen, phosphorus, potassium, pH and electrical conductivity are used to sketch the relationship among these parameters and predict the values of NPK. The predicted NPK data shows an accuracy of approximately 80% when compared with the actual dataset. These results improve the decision-making capabilities of farmers in applying right quantity of fertilizers and increase crop production.**

Keywords—Agriculture, soil nutrients, machine learning, multiple linear regression, NPK prediction

I. INTRODUCTION

In soil, nutrients like nitrogen, phosphorus and potassium are the major ones which play an important role in crop growth. In order to know these soil nutrient levels, laboratory test is carried out. Since it is a time-consuming process, farmers are applying inadequate quantity of fertilizers to achieve their yield by skipping the lab tests. Site specific soil information is required for sustainable plant growth, farm and crop management. Soil analysis consists/ of measuring various parameters namely soil moisture, nutrients, pH, Electrical Conductivity (EC), temperature and humidity. Testing the soil frequently helps farmers to know about the soil health status and this results in the improvement of their decision-making capabilities [1].

Soil testing in laboratory involves several stages like collection of soil samples at different places, mixing with right quantity of chemical substances and reagents for nutrient determination, calculating the nutrient content in the sample and finally providing recommendations based on the analysis of the existing nutrients. The nutrient content may vary according to the type of crop cultivated, change in weather and soil parameters.

Fertilizer is a key component in agriculture and precise application of fertilizers is advantageous as it improves the soil structure, increases the ability of soil water retention, stimulates healthy root development and reduces cost [2,3]. Nowadays, technology driven smart agriculture is increasing as it saves time, reduces manpower and increases yield. Machine Learning (ML) based precision agriculture manages crop efficiently and contributes to sustainable production [4]. It uses cognitive solutions to handle the issues faced by the farmers in decision-making. Various classification and regression algorithms are used to classify and predict different agricultural parameters. ML and Internet of Things (IoT) technologies assist in finding the correlation, prediction and estimation of field data and automates the agricultural techniques [5]. Hence our proposed method estimates soil NPK using a ML algorithm.

II. RELATED WORKS

The paper proposed by Ahmed et al. [6] is about Interactive Genetic Algorithm (IGA) model. Generally, macro-nutrients are analyzed in multiple approaches. The Interactive Genetic Algorithm model uses time-series sensor data and recommends fertilization based on the type of crop in a seasonal manner. The paper proposed by Sirsat et al. [7] is about an automatic system which is used to predict nutrients in the soil using regression methods. An automatic village wise fertility index prediction system is deployed in Maharashtra that predicted four nutrients namely phosphorus pentoxide, zinc, iron and manganese using seventy-six regression methods.

The nutrients of pistachio leaves are analyzed using a Genetic Algorithm and Artificial Neural Network (GA-ANN) hybrid model as proposed by Pourmohammadalia et al. [8]. The GA-ANN model is operated in MATLAB software which

978-1-6654-0816-5/22/$31.00 ©2022 IEEE

solved various complex and multi-dimensional problems. Li et al. [9] proposed a nutrient prediction system using three models namely Support Vector Machine (SVM), MLR and ANN. The system is designed to predict soil nutrients based on the relation between the independent variables such as organic matter, available phosphorus, total nitrogen, available potassium and alkali-hydrolysable nitrogen and the dependent variable total nutrient content.

Wu et al. [10] proposed a generalized regression neural network model along with K-Nearest Neighbor (KNN) and SVM. The model is used for soil nutrient prediction in China. The KNN approach shows high accuracy than regression neural network model. Ghosh et al. [11] built a model in which the soil properties such as essential plant nutrients, micronutrients and organic matter are analyzed using Supervised Learning and Back propagation Neural Network. Results proved that this back propagation method with certain neurons in the hidden layer showed better performance than supervised learning method.

Soil nutrient management using decision tree Iterative Dichotomiser 3 (ID3) algorithm is proposed by He et al. [12] which offers an efficient classification method for soil nutrient management zones and also guides the variable-rate fertilization. This technique is very challenging and requires more computation time. Hakkim et al. [13] proposed an intelligent system using MLR. The system predicts the soil nutrients and the data is sent to the low-cost fertigation system for automatic addition of fertilizers. Sreehari and Satyajee Srivastava [14] developed a model to predict rainfall conditions which helps the farmers to protect their crops. The test results shows that MLR model is far better than Single Linear Regression (SLR) model. Hence a soil nutrient prediction system is proposed using MLR method that predicts the target NPK values. This method predicts the nutrients at a good rate and it acts as a solution for the conventional time-consuming laboratory method.

III. PROPOSED SYSTEM

The proposed model uses ML algorithm named MLR to predict the NPK values based on the trained dataset. MLR is an extension of linear regression model as it combines more than one independent variable to predict the dependent variable. MLR is an important regression algorithm where a linear relationship is obtained between all the feature variables and a single target variable. The general equation of multiple linear regression with four independent variables is represented in (1). The accuracy of the model is determined by calculating R2 score, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as given in (2), (3) and (4).

$$\mathrm{Y} = \mathrm{b}_0 + (\mathrm{b}_1\mathrm{X}_1) + (\mathrm{b}_2\mathrm{X}_2) + (\mathrm{b}_3\mathrm{X}_3) + (\mathrm{b}_4\mathrm{X}_4) + \mathrm{e} \qquad (1)$$

Where,

Y is the target or dependent variable

X_1, X_2, X_3, X_4 are the predictor or independent variables

b_0 is the intercept of the regression line

b_1, b_2, b_3, b_4 are the slope of the regression lines

e is the error

$$R2 = 1 - \frac{RSS}{TSS} \qquad (2)$$

Where,

RSS- Sum squared regression

TSS – Total sum of squares

The sum of squares due to regression signifies how well the data is represented by the regression model. The variation in the observed data is measured using the total sum of squares.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (3)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (4)$$

Where,

n is the number of observations

y_i is the actual values obtained

$\hat{y}_i$ is the values determined by the model

The system is divided into two phases namely training phase and testing phase as shown in fig. 1. In training phase, the soil nutrient dataset is collected and the values are stored in excel sheet and converted into csv format. The soil dataset is taken from Kaggle repository consisting of N, P, K, pH and EC soil parameters. Python programming language is used to predict the NPK values using MLR technique, and it is also used for constructing various graphs for analysis. In this model, the soil nutrient dataset is read using pandas library and the independent and dependent variables are defined. The model is trained using MLR technique and the NPK values are predicted. In the testing phase, the predicted value is compared with the actual value and the deviation in the result is viewed with the help of scatter plot using matplotlib library.

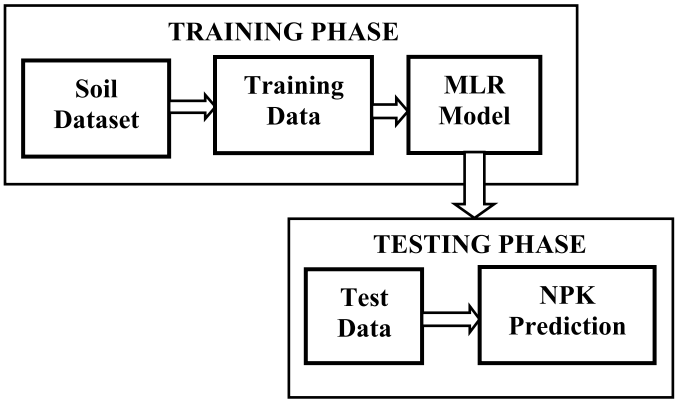

Fig. 1. MLR Model for NPK Prediction

Fig. 2 shows the MLR process workflow where the soil nutrient dataset is imported and the regression equation is applied for predicting the NPK values. The strength of the association between the predictor and target variables are estimated by using the regression coefficient formula.

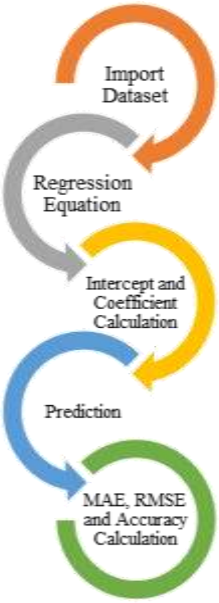
Import Dataset
Regression Equation
Intercept and Coefficient Calculation
Prediction
MAE, RMSE and Accuracy Calculation

Fig. 2. MLR Process Workflow

The intercept and the co-efficient are obtained and the slope is calculated accordingly. The performance of the system is measured by calculating the values of RMSE, MAE and R2 score. The MAE is the average magnitude of errors and it is calculated by adding the absolute value of the residuals and then dividing it by the total number of observations taken. The RMSE is obtained by finding the square root of the variance of the error obtained between a predicted value and actual value. This indicates the fineness in the predicted data values to the model's predicted data points. R2 score indicates the proportionality between the data values that lie within the regression line. It ranges from zero to one. A higher value in R2 score indicates high accuracy in prediction. The histograms are plotted for RMSE, MAE and R2 score respectively for predicting the soil dataset.

A. *Algorithm: Multiple Linear Regression model for NPK Prediction*

```
1.  Begin
2.  Read N, P, K, pH, EC
3.  Import libs
4.  Find distribution frequency of N, P, K, pH, EC
5.  x ← Independent variables
6.  y ← Dependent variable
7.  Split x_train, x_test, y_train, y_test
8.  mlr ← LinearRegression()
9.  mlr.fit(x_train, y_train)
10. y_pred_mlr = mlr.predict(x_test)
11. Predict N, P, K
12. MAE = metrics.mean_absolute_error(y_test,
    y_pred_mlr)
13. R2 = r2_score(y_test,y_pred_mlr)
14. Visualize the model using plot()
15. End
```

The algorithm begins by loading the dataset into the system, using MLR model for prediction, classifying the train and test data and computing the resultant NPK values.

IV. RESULTS

MLR is used to provide a mathematical relationship between several variables. Since only a single parameter is foreseen on the basis of many other parameters, the MLR model is implemented in this work and it is also a simple and efficient technique. This paper contributes a ML based prediction analysis method to the farmers for effective nutrient assessment and crop production.

The soil dataset consisting of N, P, K, pH and EC variables are taken and the NPK values are predicted using MLR model. To understand the distribution of these variables, histograms are plotted as shown in figs. 3 and 4. The data is grouped into bins and the bar represents number of observations or frequency per bin. These histograms help in understanding the data, discovering patterns and analyzing the soil characteristics effectively.

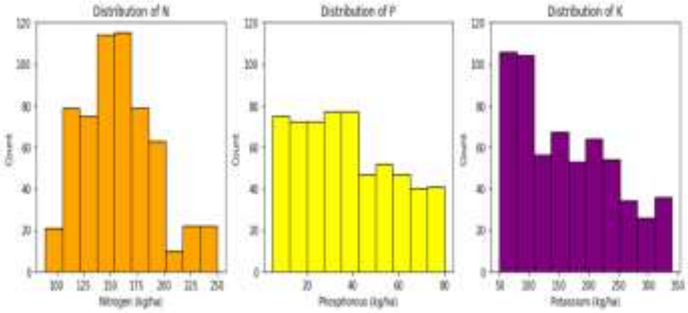

Fig. 3. Distribution of Histograms for the Variables N, P, K

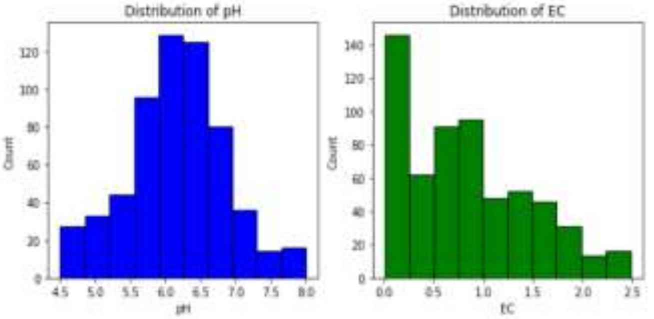

Fig. 4. Distribution of Histograms for the Variables pH, EC

The independent and dependent variables are classified for NPK prediction and the dataset is divided into 90% for training and 10% for testing. The MLR model is applied to the training data and finally the prediction for the test data is made. The actual and the predicted value is obtained and the accuracy of the MLR model is calculated. The predicted NPK values are compared with the traditional values of NPK which ranges in three scales namely low, medium and high as shown in table 1. The results obtained from MLR method can be compared by the farmers with the standard ranges of NPK and the fertilizers can be applied accordingly.

TABLE 1. STANDARD VALUES OF NPK

Nutrient (kg/ha)	Low	Medium	High
Nitrogen (N)	Below 280	280-450	Above 450
Phosphorus (P)	Below 11	11-22	Above 22
Potassium (K)	Below 120	120-280	Above 280

The MLR model shows an accuracy of approximately 80% for NPK prediction. The MAE and RMSE is calculated and it shows minimal error for the predicted NPK values. Fig. 5 represents the accuracy for all the three predicted nutrients.

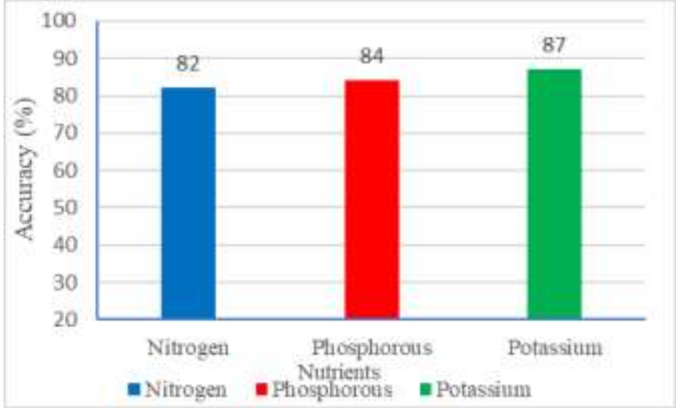

Fig. 5. Accuracy of the Predicted MLR Model

The graphs are plotted for the predicted values of NPK with all the independent variables. MLR with best R2 score has to fit the line through multidimensional data points in the plot. Most of the values predicted in this system lie in the linear regression line that best fits the data resulting in good correlation and reduced errors. The scatter plots for MLR model based on independent and dependent variables for N, P and K prediction are shown in figs. 6, 7 and 8.

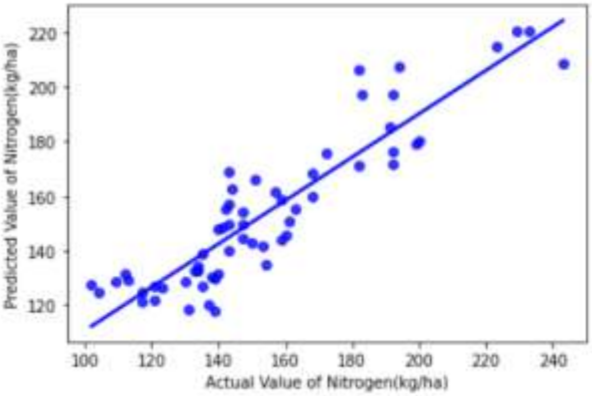

Fig. 6 Actual vs. Predicted Nitrogen Values

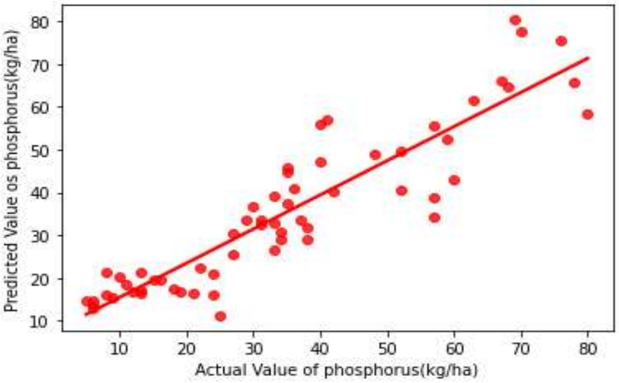

Fig. 7. Actual vs. Predicted Phosphorus Values

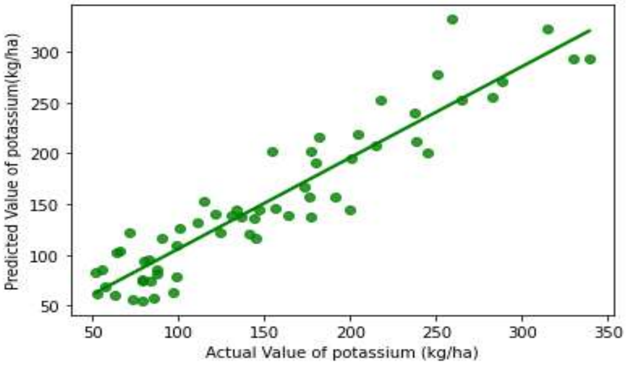

Fig. 8 Actual vs. Predicted Potassium Values

V. CONCLUSION AND FUTURE WORK

ML is employed in agriculture to improve the field performance of the crop, predict crop yield and nutrients available, detect diseases and maintain its health. ML based precision agriculture transforms farms to smart farmlands and it effectively utilizes crop inputs namely irrigation water, fertilizers, tillage and pesticides to yield high productivity without degrading the environment. The prediction of NPK values based on the independent and target variable gives an accuracy of approximately 80%. The error metrics such as MAE and RMSE shows minimal value for the proposed model. This proposed model will be useful to know the macro nutrient levels in the soil. This cost-effective system could improve the decision-making capabilities of the farmers, optimize the fertilizer application and increase the crop productivity. The future scope of this system is to predict the micronutrients like iron, zinc, sulphur etc using MLR technique.

ACKNOWLEDGMENT

The authors are grateful to the Science and Engineering Research Board (SERB) of DST for the required financial support and the Director, ICAR-SBI, Coimbatore and the Management and Principal of Sri Ramakrishna Engineering College, Coimbatore for extending the facilities for this work.

REFERENCES

[1] A. Chitragar, S. M. Vasi, N. Sujata, J. K. Akshata and I. H. Taradevi, “Nutrients detection in the soil”, *International Journal on Emerging Technologies (Special Issue on ICRIET),* vol. 7, no. 2, pp. 257- 260, 2016.

[2] T. R. Peck and S. W. Melsted, “Field sampling for soil testing”, *Soil Testing and Plant Analysis: Part I Soil Testing*, vol. 2, pp. 25-35, 1967.

[3] S. Assefa and S. Tadesse, “The principal role of organic fertilizer on soil properties and agricultural productivity - A Review”, *Agricultural Research and Technology*, vol. 22, no. 2, 2019.

[4] R. Bongiovanni and J. Lowenberg-Deboer, “Precision agriculture and sustainability”, *Precision Agriculture*, vol. 5, pp. 359-387, 2004.

[5] R. Madhumathi, T. Arumuganathan and R. Shruthi, “Internet of Things in precision agriculture: A survey on sensing mechanisms, potential applications and challenges”, in *Intelligent Sustainable Systems, Lecture Notes in Networks and Systems*, vol 213, pp. 539 – 553, 2021.

[6] U. Ahmed, J. C-W. Lin, G. Srivastava and Y. Djenouri, “A nutrient recommendation system for soil fertilization based on evolutionary computation”, *Computers and Electronics in Agriculture*, vol. 189, 2021.

[7] M. S. Sirsat, E. Cernadas, M. Fernández-Delgado and S. Barro, “Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods”, *Computers and Electronics in Agriculture*, vol. 154, pp. 120-133, 2018.

[8] B. Pourmohammadalia, M. H. Salehia, S. J. Hosseinifardb, I. E. Boroujenic and H. Shiranic, “Studying the relationships between nutrients in pistachio leaves and its yield using hybrid GA-ANN model-based feature selection”, *Computers and Electronics in Agriculture*, vol. 172, pp. 1-7, 2020.

[9] H. Li, W. Leng, Y. Zhou, F. Chen, Z. Xiu, “Evaluation Models for Soil Nutrient Based on Support Vector Machine and Artificial Neural Networks”, *The Scientific World Journal*, vol. 2014, pp. 1-7, 2014.

[10] C. Wu, Y. Chen, X. Hong, Z. Liu and C. Peng, “Evaluating soil nutrients of Dacrydium pectinatum in China using machine learning techniques”, *Forest Ecosystem*, vol. 7, no. 30, 2020.

[11] S. Ghosh, S. Koley, “Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks”, *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 2, no. 2, pp. 292-297, 2014.

[12] L. He, C. Liying, C. Guifen and L. Dexin, “Delineating Soil Nutrient Management Zones Based on ID3 Algorithm”, in *Proc. International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)*, Jilin, China, pp. 1155-1159, 2011.

[13] A. Hakkim, A. Joseph, A. Gokul, and K. Mufeedha, “Fertigation: A Novel and Efficient Means for Fertilizer Application”, *International Journal of Current Research*, vol. 8, no. 8, pp. 35757-35759, 2016.

[14] E. Sreehari and Dr.Satyajee Srivastava, “Prediction of Climate Variable using Multiple Linear Regression”, in *Proc. 4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, pp. 1-4, 2018.