Enhancing Precision Agriculture Through Novel Soil Nutrient Prescription Recommender System Using Quantile Regression Forests

Sareena Rose[1], S Nickolas[2], S Sangeetha[2]

Vimala College (Autonomous), Thrissur, Kerala, India[1]

Department of Computer Applications, NIT Trichy, Tamil Nadu, India[2]

Corresponding author: Sareena Rose, Email: sareenarose@vimalacollege.edu.in

The conception of a soil nutrient-specific prescription recommender system will streamline the understanding of the cause-and-effect relationships inherent to diverse nutrients and help in informed decision-making about farming methodologies. Prescriptive analytics proposes optimal solutions Using Quantile Regression Forests and provides recommendations that enable soil scientists to understand the sufficiency/deficiency of soil nutrients at customised locations. Unlike traditional methods that rely on spatial orientations, the proposed approach focuses on the proximity between data points by prioritizing the influence of regressor effects over spatial factors. The crux of this paper lies in the prescription and recommendations, which were derived using a customized algorithm, centered around the proximal data points. The proposed system exhibited notable accuracy rates when compared with three alternative approaches (kNN, Decision Tree, Naive Bayes) with accuracy rates of 97.4% and 89.6% in providing recommendations for response and input nutrients, respectively.

Keywords: Quantile Random Forests, Recommender systems, Prescriptive analytics, Proximity matrix

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 657–675. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-59

1 Introduction

Soil, the earth's skin, is the repository of various nutrients and the make-or-break factor in agriculture. It is integral in protecting the earth's surface and is greatly influenced by geographical, environmental and weather parameters. Its rich nutrient and mineral content plays an eminent role in regulating the essence of the ecosystem. Any malfunction of the soil properties affects not only agriculture but also the water cycle of the earth and, to an extent, paves the way for natural calamities. Hence, it is a thriving need today to sustain soil's biological and chemical composition as it is.

The assessment of micronutrients in a soil sample is a multivariate problem modelled as a combination of the soil's macronutrients and their physical and chemical properties. Despite these, the chemical compounds of the nutrients may change according to time, geographical and climatic conditions. These correlations, among many factors, contribute to the soil's heteroscedastic nature and force it to consider the samples' locality while processing the data. So, under the constraints of non-linear relationships among the variables and their heteroscedastic nature, the assessment of essential micronutrients is a cumbersome procedure that cannot be generalized. To improve the accuracy of the soil nutrient prediction, it is better to have a customised model for each geographic region which is practically difficult.

The macronutrients of the soil can be assessed in an economically feasible and at a faster pace. These variables serve as the input to predict the micronutrients and bring precision to agriculture. While dealing with such predictions, the impact of leverages among the input attributes is an important factor that cannot be overlooked. In the case of soil samples, there is no generalised trend to understand what is wrong with the leverage points; instead, the locality of the samples needs to be investigated. Another bottleneck is the correct interpretation of the sufficiency or deficiency of the assessed micronutrients. To have an accurate judgement the soil scientists need to know the geographical characteristics, soil type and the other parameters that have influenced the soil formation. This approach is not practically viable; hence, soil scientists have a generalised chart to determine the sufficient range of nutrient values among heterogeneous soil samples. Together, soil nutrient prediction and recommendation systems require the statistical and machine learning techniques to be intertwined to address the heterogeneity and heteroscedasticity of the soil samples.

A recommendation system is a subclass of information filtering systems that can foresee user preferences using Data Science [1]. There are three types of rec-

ommender systems based on the user and item-based approaches. The Content-based recommendation systems focus on the user's preferences and learn all the features liked and disliked by the user. The knowledge learnt through this method is used in future recommendations. In contrast, collaborative filtering-based systems learn the similarities among the users rather than items liked by the users to make a future recommendation. These systems use the neighbourhood of the test user to understand his preferences, and this neighbourhood is learnt using the similarity finding algorithms in Machine Learning such as k-NN, Bayesian theorems and Decision Trees. The third approach is hybrid recommender systems which work in a hybrid mode, using both content and collaborative approaches.

Plenty of research is happening on implementing recommender systems in the agricultural sector. The data from various resources like sensors, satellites and databases are integrated, cleaned and processed to gain useful insights and to provide recommendations. A schematic representation of recommender systems in agriculture is shown in Fig. 1. The state-of-the-art recommendation systems are mainly used to recommend crops, fertilisers and soil quality. But, a prescription-cum-recommendation system is a novel idea in the agricultural sector, especially in checking soil quality. In collaborative recommender systems, a similarity matrix is constructed to assess the proximity among the subjects under study. This matrix is built using ML techniques which examine their neighbourhood using various distance metrics. In the proposed recommender system, the effect of the regressors on the predicted variable needs to be examined to build the similarity matrix rather than the spatial orientation. This is a challenge with respect to any heterogeneous and heteroscedastic datasets. Hence, the proposed system recommends using the Quantile Regression Forests to assess the soil samples' similarity using a proximity matrix.

This article focuses on developing a simple customised recommender system that identifies whether the soil is sufficient or deficient in nutrients and recommends it for cultivation. When an unseen observation comes in, the system checks the inputs and undergoes the prediction process only if it is compatible with the regular points. In the opposite case, a diagnosis must be made, and a prescription must be given to the stakeholders that suggest the most suitable values. The relevance of this recommender system is to help farmers to make wise and judicious decisions about crop cultivation knowing the nutrient content of their soil. Such a recommender system can replace the conventional methods of analysing the soil samples to learn the similarities and differences to recommend the nutrients they lack. The prerequisites to build this recommender system is labelled sets

of regular and anomalous observations and efficient mechanism to predict the micronutrients.

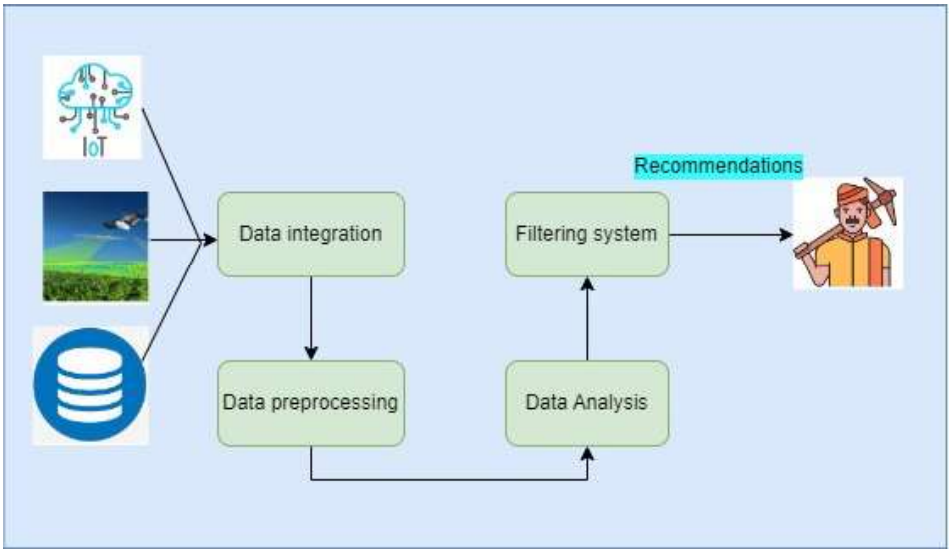

Figure 1: A schematic illustration of recommender systems in agriculture

2 Literature Survey

The primary objective of a Recommendation System (RS) is to foresee the needs and interests of the user and recommend the appropriate entities/items/services. The research on recommender systems dates back to early 90ś when the first recommender system based on collaborative filtering is proposed [2]. Later, plenty of research was carried out and a strong foundation on the theoretical concepts and practical applications of the recommender systems are proposed in [3], [4], [5], [6], [7]. The implications and challenges of recommender systems are also investigated and remedies and optimal solution for each one is also studied [9], [10], [11], [12].

Prescriptive analytics, a forerunner of the recommender system, proposes optimal solutions based on the data analysed leading to providing actionable insights and recommendations for decision-making. An extensive review of the field of prescriptive analytics and the importance of optimization techniques, decision-making models and their applications in various industries is highlighted in [13].

A comprehensive survey focusing on the utilization of prescriptive analytics to demonstrate its efficacy within the realm of big data analytics is seen in [14]. Kandula et al. [15] propose a prescriptive analytics framework tailored to enhance the efficiency of E-commerce order delivery processes. The authors make use of delivered order's status and an algorithm is designed to generate order success profiles which are further optimized for decision-making procedures. Srinivas et al. [16] attempt to optimize the out-patient appointment system by categorizing patients as high-risk and low-risk types and depending on their type of visits, the appointments are scheduled. Yange et al. [17] use prescriptive analytics to find the optimal crop yield using Extreme gradient Boosting and SVM in their research to improve the profit of the farmers.

The agriculture domain has been bestowed with numerous studies and research on recommender systems rather than prescriptive analytics. The various works on recommender systems include suggesting the suitability of land, crops to be cultivated, pesticides to be used etc. Pudumalar et al. [18] proposed a crop recommender system based on the geographical and weather properties and in-site properties using Random Forests. A similar type of recommender system can be seen in the [19] using Neural Networks. Pande et al. [20], on getting the location of the land and the soil type as the input, recommend the crops to be cultivated. The amount of fertilisers is recommended based on the quantity of macronutrients in the fuzzy logic-based recommender system proposed in [21]. Using a collaborative recommender system, the recent trends in the agricultural field and government schemes for farmers are recommended by Jaiswal et al. [1]. In [8] recommends the top five crops based on the soil nutrient status; here, they use the inputs available from the soil health card of each land. On identifying the diseases affecting the plants, proper nourishment through fertilisers is recommended in [22]. Senapaty et al. [23] used the Multi-class Support Vector Machines to accurately classify the soil nutrients and suggest corresponding crops. In all of the above systems, different machine learning models are employed and they produce better accuracy in the recommendations suggested.

2.1 Research gaps and contributions

The research gaps identified in a few of the above systems are that they rely on primary/ secondary output data to provide recommendations, whereas, the rest provide recommendations on the outputs after predicting that from the inputs. The former systems have a higher probability of error if inputs provided to them are inaccurate. The latter eradicates such scenarios; however, these system's recommendations are based on various customised optimization techniques. If the optimisation techniques fail, the credibility of these suggestions stands questionable. The proposed method stands out from the existing scenario by predicting the nutrients using the primary data and providing recommendations based on the model built during the dat cleaning eliminating the need for any other optimization techniques. As long as the model built for prediction is fine-tuned, then the recommendation engine will also work well. Another contribution of the proposed system is to provide recommendations with respect to the leverages among the input values. This is a novel idea as far as to the best of our knowledge as recommendations are only provided to the response variables and any error in the predictor variables is ignored.

3 Methodology

3.1 Quantile Regression Forests (QRF)

Like Random Forest (RF), trees are grown in QRF, but they differ in how the leaves store the information. RF stores the averages of all observations in the leaves, satisfying the nodes criterion, whereas QRF stores the observations [24]. This property helps QRF achieve conditional estimates at various quantiles of the entire distribution and is helpful in a heteroscedastic context for an accurate extrapolation. Examining the quantiles will unveil many exciting patterns of data and the quantified impact of predictors on the response variable, which throws light into the unusual combinations of the predictors. To build a QRF, there are two significant hyperparameters in RF which need fine-tuning: the minimum number of response values stored in a leaf and the number of trees. As advised in [25], the leaf size should be a minimum of five or higher, as a lesser value tends to attract noise. Oshiro et al. [26] advocate that merely increasing the number of trees in a Random Forest does not make the forest efficient in prediction but makes it computationally inefficient. The study of error estimates in [25] shows that out-of-bag (OOB) error estimates are worthy of being considered as a validation strategy to decide on the efficiency of Random Forests and hence are used

here to estimate the best hyperparameter values. The optimisation parameters to train the model are chosen after an exhaustive test strategy. To determine the leaf size, values are assumed from 5 to 10% of the size of the data set and the cumulative OOB error for all trees is calculated. The leaf size with the minor OOB error is selected as the optimal value for that parameter. The values ranging from 100 to 50% of the size of the data set are tested to choose the best ensemble size, and the one with the lowest cumulative OOB error is selected as the number of trees in this algorithm.

Koenker [27] states that "segmenting the samples into subsets defined according to conditioning covariates is a valid option" in quantile regression. When a multivariate random forest is used as a tool for quantile regression, the conditioning of covariates is implicitly implemented by injecting randomness at two levels while expanding the nodes of the tree and while selecting the subsamples for the construction of each weak learner [28]. This enhances the prediction accuracy of the quantiles. Here random subset of predictors is decided using the interaction-curvature test, and the split points are determined by taking the impurity levels of the resultant nodes.

3.2 Classification of observations using proximity matrix

In heteroscedastic datasets, the similarity is measured in terms of the predictors' regression effect, unlike the state-of-the-art distance measurements focusing on spatial orientation. This is one significant aspect of this algorithm, and it uses the proximity matrix of the RF for the same. This matrix is an N-by-N matrix, where N is the number of observations, its diagonal elements are 1, and it is a symmetric matrix ($a_{ij} = a_{ji}$). In RF, if two observations fall into the same leaf node, they are treated as homogeneous [28] and this property is used here to identify the regular observations from the leverage points. The proximity matrix is computed in the following manner. The observations i and j are run down through each tree in the RF, and each time both observations fall into the same leaf node, their count is incremented by 1. After finding the common trees in the forest for i and j, the count is normalised by dividing it by the total number of trees. The proximity value ranges from 0 to 1 for any observation in the dataset and is crucial in determining the status of the observations.

3.3 Check the eligibility for the prediction

The unseen test point is run down through the trees of QRF, and the first proximal neighbour n is chosen. If at least half of the trees in the forest agree with x and n, then x belongs to the group of n. If there is no such n, then the majority voting is performed on the neighbours who agree with x in at least 10% of trees and x is assigned to that group of the major neighbour. If the observation is fit for prediction, it is fed into the system for prediction. If there is a leverage point in the variables, they are treated as leverage entries and are not allowed for prediction. Instead, their nearest neighbour is voted using majority ranking and prescribed as the expected value for the leverage point .

3.4 Prediction system and recommendations for regular points

When an unseen data point comes and if it is a regular point, the prediction system predicts all micronutrients quantitatively. The IF-Then rules developed with the inputs from soil experts are applied to these quantitative measurements of the nutrients, and recommendations are provided. The recommendations include whether the soil sample has sufficient nutrients for cultivation, and if there is a deficiency, the system prescribes the most suitable value required. The proximal data point, sufficient in the micronutrients, is found using QRF and recommended as a remedy for this deficiency.

3.5 Recommendations for leverage points

Once the leverage point is identified, its most proximal data point, which is a regular observation, is estimated using the QRF trees. This proximal point's values are provided as a recommendation to the stakeholders to have a good idea of the expected value for that soil sample. Let R be the set of reference points and j be the test point; then, $j's$ proximity with R is computed. And the most proximal reference point is chosen, and its proximity is recorded. The data point is a global outlier and discarded if there is no such proximal point. Otherwise, the vectors $Z1$ and $Z2$ are computed that represent the minimum and maximum predictor values of R. This serves as the boundary for the normalisation process and is utilised to normalise the pair of proximal observations in R and j. The normalised values will help to quantify and highlight the similarity among the observations in R and j.

The difference between the normalised values of R and j is calculated and penalised if the observation falls outside the allowed range. The penalty measure

Algorithm 1 Find_Leverage (R,wt)

Input:R is the set of regular points, j is the leverage obs., wt is the weight of the predictors

Output:j_n, **the transformed point**, θ, **the indices of the transformed predictors**

Procedure

Z1=min(R)
Z2=max(R)
P= proximal referral point to j in R
If P is not NULL
    Repeat
        N_s=Normalise j w.r.t $Z1$ and $Z2$
        N_r=Normalise P w.r.t $Z1$ and $Z2$
        ϕ= penalty calculated based on N_s and N_r
        N_{diff}=abs(N_s - N_r)+ ϕ
        N_{wt}=N_{diff} * wt
        Find the predictor p with maximum weight in N_{wt}
        Replace p's value with corresponding variable's value in P
        Record p in θ
        Check the inlyingness of the newly ceated vector j_n
    Until j_n is added to R
End

depends on how far they are off the range and is calculated using Eqn. 59.1. This penalised difference ϕ is multiplied by the standardised feature weights wt to add weightage to the significant features in prediction. Now N_{wt} represents the vector with the variability level of the predictors. The most significant variation and its predictor p are chosen. The p's value is replaced with the corresponding predictor in P resulting in a new data point j_n. This new point is subjected to the test of the measure of inlyingness. If the point remains a leverage point, the above process is repeated, and the next predictor is chosen, which varies considerably until the point transforms into a regular observation. Each change is recorded in a vector θ to prepare the prescriptions for the farmers. The whole procedure is summarized in Algorithm 1.

The normalisation technique used here is min-max normalisation, and the penalty is calculated by finding the difference between the minimum/ maximum

values of S and R's normalised vectors. Each predictor i of S and its proximal point in R is normalised w.r.t the min. and max. values of the i^{th} predictor in R. The difference between the minimum and maximum values of these normalised vectors attributes as the penalty. Suppose the queried observation and its proximal referral point are within the range of the inliers. Their normalised vectors will be $m < v < x$, where v is the normalised values of the respective observations, and m and x are the min. and max values. If the predictor falls below the range, the normalised vector will be $m > v < x$ and $v = 0$; if the predictor falls outside the range, the vector will take the form $m < v > x$ and $v = 1$. In the former case, the penalty will be the sum of m values of normalised vectors of S and its proximal vector R. In the latter case, the penalty is the sum of $(1 - x)$ of both vectors. An exceptional case is when the normalised vector of S is in the range, and R goes out of the range. Here, S is penalised because, by any chance, R is a false positive; its effect should not be favourably propagated to S. This will ensure that S passes the inliers test only if it is a true positive. The calculation of the penalty p is given as in Eqn. 59.1.

$$\phi = \begin{Bmatrix} 0, & if \quad v1 \quad and \quad v2 \quad falls \quad within \quad the \quad range \\ m1 + m2, & if \quad v1 \quad or/and \quad v2 \quad falls \quad below \quad the \quad range \\ (1 - x1) + (1 - x2), & if \quad v1 \quad or/and \quad v2 \quad falls \quad above \quad the \quad range \end{Bmatrix} \tag{59.1}$$

The recommendations are proposed in the following steps and the schematic representation of the workflow of this phase is given in Fig. 2.

4 Results and Discussion

Three secondary datasets which differ in their natural characteristics are chosen from Thrissur district, India for this study. These datasets are obtained from Ministry of Agriculture, Govt. of Kerala, India, after obtaining written consent from the concerned ministry. This data is available to the public on the government website *www.soilhealth.dac.gov.in.* The samples of Dataset 1 are collected from a midland region located at Chalakudy, $10.3070^{\circ}N$ and $76.3341^{\circ}E$, and has 6648 observations. Dataset 2 has 3618 observations and belongs to backwaters region at Kodungallur ($10.2277^{\circ}N$, $76.1971^{\circ}E$). Dataset 3 consists of samples (8018) collected from the coastal area (Chavakkad) located at 10.5782°N, 76.0191° E. These are raw data with impurities and missing values. Dataset 1 has gravelly laterite soil, which is strongly acidic and has low water retention capacity. This type of soil has good physical properties and high content of Phosphorous but is defi-

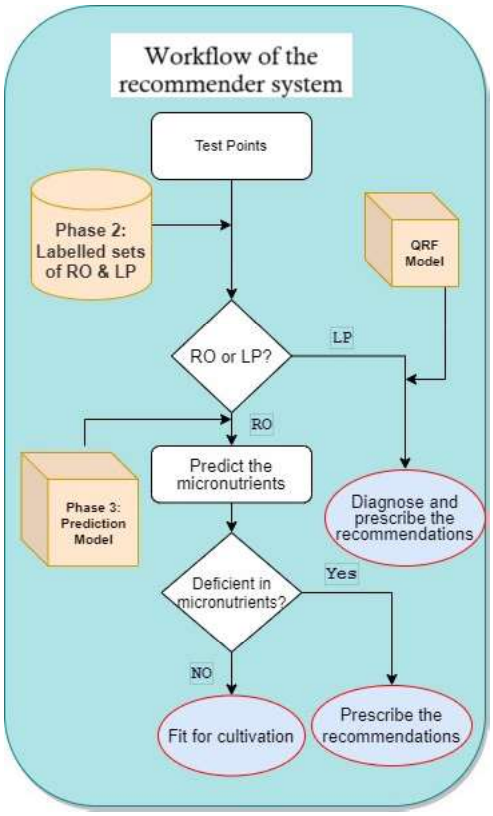

Figure 2: The workflow of Phase 4

cient in Magnesium and Boron. The samples comprising Dataset 2 have coastal sandy soils that are moderately acidic, low in organic matter, high in Phosphorous and deficient in plant nutrients. The samples of Dataset 3 have a mixture of coastal sandy soils and coastal alluvium soils. All three datasets are split in the ratio 67:33 for training and testing purposes.

For each dataset, there are five predictor variables, the trio of macronutrients (Nitrogen, Phosphorus, and Potassium), the pH value and Electrical Conductivity (EC) of the soil and six response variables. These are the micronutrients, viz, Zinc (Zn), Ferrous (Fe), Copper (Cu), Manganese (Mn), Boron (B) and Sulphur(S). The labelled sets of regular and leverage observations of each dataset are obtained as the input and the hyperparameters are optimised to build the QRF.

The minimum leaf size and the maximum number of trees are estimated for each dataset using the exhaustive test strategy specified in Section 3.1. Models of Ran-

dom Forests with bagging are then built using these optimised hyperparameters (Table 1). From the QRF model, the proximity matrix is obtained. Once the la-

Table 1: Optimized hyperparameters and the count of unique inliers, outliers and ambiguous observations identified by the soil experts for each dataset

Datasets	No: of trees	Leaf size	Regular Points	Leverage Points
DS1	100	5	6391	257
DS2	50	5	3522	96
DS3	300	5	7434	584

belled set of inliers and leverage are identified, the unseen test point is run down through the tree of QRF, and the first proximal neighbour n is chosen. If at least half of the trees in the forest agree with x and n, then x belongs to the group of n. If there is no such n, then the majority voting is performed on the neighbours who agree with x in at least 10% of trees and x is assigned to that group of the major neighbours.

If the unseen test point is an inlier it is fed into the prediction system and

Table 2: Recommendations for each nutrient for regular points that are suitable for cultivation

Datasets	P	pH	K	N	Zn	Fe	Cu	Mn	B	S
DS1	36 (S)	4.9 (HAc)	400 (S)	0.72 (S)	3.81 (S)	159.8 (S)	2.0 (S)	32.1 (S)	0.52 (S)	9.87 (S)
DS1	36 (S)	5.0 (HAc)	360 (S)	0.82 (S)	7.04 (S)	164.6 (S)	3.81 (S)	48.4 (S)	0.53 (S)	11.6 (S)
DS2	36 (S)	6.1 (MAc)	144 (M)	0.54 (S)	2.68 (S)	124 (S)	2.07 (S)	28.2 (S)	0.52 (S)	10.6 (S)
DS2	36 (S)	6.1 (MAc)	341 (S)	5.6 (H)	2.81 (S)	172.6 (S)	3.862 (S)	42.6 (S)	0.61 (S)	16.1 (S)
DS3	36 (S)	5.1 (HAc)	275 (M)	0.69 (S)	4.1 (S)	172.1 (S)	4.21 (S)	28.6 (S)	0.51 (S)	14.3 (S)
DS3	89.1 (H)	5.6 (HAc)	275 (M)	0.79 (S)	3.83 (S)	94.2 (S)	1.6 (M)	33.4 (S)	1.01 (S)	15.6 (S)

their nutrient quantity is estimated using the Gaussian Process incorporated with RAM [29]. Then the predicted nutrients are classified in terms of sufficiency and deficiency. Here, the categories are learnt from the soil experts' inputs and the proximal data points nutrient values obtained using QRF which is incoroporated into the recommender system using simple IF-Then rules. Table 2 shows one such output where all the entries are considered suitable and fit for cultivation. The characters inside the brackets are individual classifications of each nutrient, where *S stands for sufficient, H for High, M for Medium, and Ac for Acidic.* It can be seen that for Dataset 3, the last entry has a high value for P when compared to the rest, but it is still fit for cultivation. Similarly, Dataset 2 has high Nitrogen content in its soil sample but validated with the rest of the nutrient content, it is still fit for cultivation. The table also shows the heterogeneity of the soil samples

in terms of nutrient values even though the samples are collected from a single district. This is owing to the soil's dependence on geographical parameters and accounts for the need for customisable models for prediction for each region.
Table 3 shows the sample recommendations if a deficiency in the micronutrients in the soil sample is reported. Such an output is generated when the test point is an inlier but the predicted micronutrients had anomalous values. This is considered an outlier in the response variable, and the recommendations are suggested. The recommendations are based on the most proximal regular observation with a sufficient value in the same response variable, which is currently deficient in the test point. However, this recommendation only suggests the expected value a regular observation should possess. Also, in addition to the recommendation, the prescription for the suitable values of the predictors is suggested, so that the soil scientists can be clear on what grounds the deficiency could have been reported. For eg., in Dataset 1, Sulphur is found deficient in the soil sample. Hence the proximal data point of this sample, in terms of its regressors is obtained using QRF and the possible value for Sulphur is recommended. Another contribution of the proposed system is prescribing the suggested values that the predictors should have to have such a reasonable value of the nutrient. The proposed algorithm works for every datasets in the similar manner, and in order to avoid redundancy only a few is shown in the table.
The next scenario is when the test point is a leverage point. Here, the test point

Table 3: Prescriptions for the deficiency of the micronutrients for regular points

Datasets	P	pH	K	N	Zn	Fe	Cu	Mn	B	S
DS1	36 (S)	5.8 (MAc)	242 (S)	1.47 (S)	11.4 (S)	128.8 (S)	2.8 (S)	48.41 (S)	0.52 (S)	5.27 (D)
	pH and Macronutrients are suitable				Micronutrients suitable for cultivation **S is deficient**. Suitable value: **10.8** Predictors pH, K, N: [5.8 231 1.34]					
DS2	36 (S)	5.6 (MAc)	220 (M)	1.08 (S)	2.45 (S)	128.6 (S)	2.1 (S)	74.6 (S)	0.09 (D)	39.2 (S)
	pH and Macronutrients are suitable				Micronutrients suitable for cultivation **B is deficient**. Suitable value: **0.61** Predictors P, pH, K, N: [36 6.1 264 0.74]					

is not fit for prediction, the system tries to diagnose and prescribe the necessary recommendations. The prescription is prepared after gaining insights about the most proximal data point of the leverage point which is a regular point. This is obtained using the proximity matrix of QRF and the prescriptions contain the recommendations provided to the leverage point with respect to the predictors involved. In the recommendations, the anomalous predictor variable is identified and its best possible value w.r.t the proximal point is prescribed. The system also

computes the measure of the proximity of the anomalous data point to the regular point. After the leverage point is substituted with the proposed recommendations, the leverage point undergoes a proximity check to reassess the degree of outlyingness. Table. 4 shows one such output. In this context, the soil sample exhibits an unusual nitrogen content, rendering it unsuitable for predicting micronutrients accurately. In the subsequent step, we identify the nearest data point, which shares a proximity of 34.54%. The nitrogen value from this neighbouring point is then recommended as a replacement. After this substitution, the proximity to the original data point is re-evaluated. In this instance, the proximity significantly increases to 70.91%, bringing all predictor variables within an acceptable range for accurate prediction. In exceptional scenarios, substituting multiple predictor variables may be necessary to enhance the observation's proximity. This approach offers the advantage of enabling soil scientists to quickly determine optimal values for anomalous data points, without delving into intricate soil type and geographical details. This eliminates the need for extensive human intervention.

To assess the effectiveness of the proposed system, we constructed proxim-

Table 4: Status after the implementation

Nutrients	Values	Recommendations & Prescriptions	New values	Status after the implementation
P	36 (S)	Macronutrients: **N is high**. Proximity with RO: **34.54%** Prescribed value: **1.42**	36(S)	Macronutrients: Suitable **Proximity: 70.91%**
N	2.2 (VH)		1.42 (S)	
K	400 (S)		400 (S)	
pH	5.6 (S)		5.6 (S)	

ity matrices using three distinct methods: k-Nearest Neighbor (kNN) with a parameter k set to 10, Decision Trees employing interaction-curvature splits, and Naive Bayes classification based on conditional probabilities. For kNN, the proximity matrix was established by calculating the Euclidean distance between input points. In the case of Decision Trees, a single tree was constructed, and proximity values akin to Quantile Regression Forests (QRF) were recorded. Naive Bayes relied on conditional probabilities to create its proximity matrix, considering the relationships between the input point and other dataset points. These proximity matrices were then integrated into the recommendation engine of the proposed system, and the accuracy of the recommendations was documented. The graph-

ical representation of the recommendation accuracy is depicted in Figure 3. The results indicate that both nearest neighbour and decision trees performed sub-optimally in constructing proximity matrices for recommendation purposes. In contrast, the proximity matrix derived from conditional probabilities proved to be more effective than the former methods. Notably, the proposed system outperformed all three approaches, particularly in providing prescriptions for optimizing data point leverage.

Figure 3: Comparison of accuracy in recommendations

5 Conclusion and Future Scope

The domain of soil chemistry is getting revolutionised, with various electronic devices and sensors taking the place of the conventional assessment method through laboratory experiments. The major breakthrough is the use of recommender systems in agriculture to provide recommendations about crops, fertilizers based on the soil content. This paper proposes a prescription cum recommender system which recommends the suitability of the soil for cultivation and prescribes the best possible values if the sample is deficient in micronutrients. The leverageness of the input variables are also identified and the prescriptions are suggested with the help of the proximal point based on the regressor effect. A Quantile Regression Forest is trained using labelled sets of regular and leverage points and using a customised algorithm the prescription cum recommendation system is implemented. The advantage of the proposed approach is that it does not rely on a range or a generalised equation; it does not need any prior knowledge yet provides a customised and reliable recommendations, that is the need of any heterogeneous and heteroscedastic dataset in ecological domains. The point of concern in this approach is the need for extensive sample size for training this model, lest the algorithm fails to provide proper suggestions. Another demerit is that the proposed algorithm prescribes only a known and preferable value not the exact suitable value and this is the scope of future work for the authors.

References

[1] Jaiswal, S., Kharade, T., Kotambe, N. and Shinde, S.: Collaborative recommendation system for agriculture sector. In: ITM web of conferences. vol. 32, pp. 03034. EDP Sciences. (2020).

[2] Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Comm. of ACM. 35(12), pp. 61-70 (1992).

[3] Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender Systems Survey. Knowl Based Syst, 46, pp. 109-132 (2013).

[4] Resnick, P. and Varian, H.R.: Recommender systems. Comm. of ACM. 40(3), p.56-58 (1997).

[5] Aggarwal, C.C.: Recommender systems. vol. 1. Cham: Springer International Publishing. (2016).

[6] Melville, P. and Sindhwani, V.: Recommender systems. Encyclopedia of Machine Learning. 1, pp. 829-838 (2010).

[7] Lü, L., Medo, M., Yeung, C.H., Zhang, Y.C., Zhang, Z.K. and Zhou, T.: Recommender systems. Phy. Repts. 519(1), pp. 1-49 (2012).

[8] Patel, K. and Patel, H.B.: A state-of-the-art survey on recommendation system and prospective extensions. Comput. Electron. Agric. 178, p.105779 (2020).

[9] Khusro, S., Ali, Z. and Ullah, I.: Recommender Systems: Issues, Challenges, and Research Opportunities. In Information Science and Applications (ICISA) 2016, Springer Singapore. pp. 1179-1189 (2016).

[10] Milano, S., Taddeo, M. and Floridi, L.: Recommender systems and their ethical challenges. AI & Society, 35, pp. 957-967 (2020).

[11] Mohamed, M.H., Khafagy, M.H., Ibrahim, M.H.: Recommender Systems Challenges and Solutions Survey. In 2019 International Conference on Innovative Trends in Computer Engineering (ITCE) pp. 149-155, IEEE (2019).

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 657–675. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-59

[12] Ricci, F., Rokach, L., Shapira, B.: Recommender Systems: Introduction and Challenges. Recommender Systems Handbook, pp.1-34 (2015).

[13] Lepenioti, K., Bousdekis, A., Apostolou, D. , Mentzas, G., Prescriptive analytics: Literature review and research challenges. Int. J. Infn. Mangt., 50, pp. 57-70 (2020).

[14] Poornima, S. and Pushpalatha, M.: A survey on various applications of prescriptive analytics. Int. J. Intt. Net., 1, pp. 76-84 (2020).

[15] Kandula, S., Krishnamoorthy, S., Roy, D.: A prescriptive analytics framework for efficient E-commerce order delivery. Decision Support Systems, 147, 113584. (2021). doi:10.1016/j.dss.2021.113584

[16] Srinivas, S., Ravindran, A. R.: Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics framework. Expert Systems with Applications, 102, 245–261 (2018). doi:10.1016/j.eswa.2018.02.022

[17] Yange, S.T., Egbunu, C.O., Rufai, M.A., Onyekwere, O., Abdulrahman, A.A., Abdulkadri A.: Using prescriptive analytics for the determination of optimal crop yield. Int J Data Sci Anal. 6(3) pp. 72-82 (2020).

[18] Pudumalar, S., Ramanujam, E., Rajashree, R.H., Kavya, C., Kiruthika, T., Nisha, J.: Crop recommendation system for precision agriculture. In 2016 Eighth International Conference on Advanced Computing (ICoAC), IEEE, pp. 32-36 (2017).

[19] Priyadharshini, A., Chakraborty, S., Kumar, A. and Pooniwala, O.R.: Intelligent Crop Recommendation System Using Machine Learning. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, pp. 843-848 (2021).

[20] Pande, S.M., Ramesh, P.K., Anmol, A., Aishwarya, B.R., Rohilla, K. and Shaurya, K.: Crop Recommender System Using Machine Learning Approach. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE pp. 1066-1071 (2021).

[21] Haban, J.J.I., Puno, J.C.V., Bandala, A.A., Billones, R.K., Dadios, E.P., Sybingco, E.: Soil Fertilizer Recommendation System using Fuzzy Logic. In 2020 IEEE REGION 10 CONFERENCE (TENCON), IEEE pp. 1171-1175 (2020).

[22] Selvi, D.P. and Poornima, P., Soil Based Fertilizer Recommendation System for Crop Disease Prediction System. Int. J. Engg.Trends Appln. 8(2) (2021).

[23] Senapaty, M. K., Ray, A., Padhy, M.: IoT-Enabled Soil Nutrient Analysis and Crop Recommendation Model for Precision Agriculture. Computers. 12(3):61 (2023). doi.org/10.3390/computers12030061

[24] Meinshausen, N. and Ridgeway, G.: Quantile Regression Forests. J. Mach. Learn. Res. 7(6) (2006).

[25] Breiman, L.: Random forests. Mach Learn, 45, pp. 5-32 (2001).

[26] Oshiro, T.M., Perez, P.S. and Baranauskas, J.A.: How many trees in a random forest?. In Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8, Springer Berlin Heidelberg pp. 154-168 (2012).

[27] Koenker, R. and Hallock, K.F.: Quantile Regression. J Eco Pers, 15(4), pp. 143-156 (2001).

[28] Segal, M. and Xiao, Y.: Multivariate Random Forests. Wiley interdisciplinary reviews: Data Mining and Knowledge Discovery, 1(1), pp. 80-87 (2011).

[29] Rose, S., Nickolas, S. and Sangeetha, S.: Effective prediction of soil micronutrients using Additive Gaussian process with RAM augmentation. Comput. Biol. Chem. 98, p.107683 (2022). doi.org/10.1016/j.compbiolchem.2022.107683