Predictive Soil Analytics Using Data Mining Techniques

Ms.G.Rubia[1]
Dr.M.Nandhini[2]

[1] Department of Computer Science, Government Arts College, Udumalpet, Tamilnadu, India-642126.
[2] Department of Computer Science, Government Arts College, Udumalpet, Tamilnadu, India-642126.
[1]grubia15@gmail.com
[2]nandumano@gmail.com

Abstract. Agriculture and allied industries play an important role in the development of our nation. In India more than 55% of people make a living from farming. Crop yields are an essential aspect of every farmerâs day. It depends on many factors like soil quality, seeds, planting practices, humidity, fertilizers and pesticides. Besides all factors, diagnosing soil quality is a fundamental and essential task in farming, as it provides background knowledge of the soil and its physical, chemical and biological prominence. Hence, soil analytics is inevitable that gives information about the present nutrient availability or the need of the nutrients for effective cultivation. It helps to interpret the physico-chemical properties of soil nutrients and to classify the nutrient content as very low, low, medium, high, or very high based on pH values. Thus, predictive analytics based on the soil parameters offer precise and sensible solutions for soil fertility problems and enable suitable decisions on crop cultivation. This study attempts to exploit the benchmark classification algorithms from data mining to classify soil samples of Tiruppur district using pH levels. The prediction of pH levels is important to know the nutrients availability in the soil. Classification algorithms like Logistic Regression (LR), Bernoulli Naive Bayes (BNB), Decision Tree (DT), Extra Tree (ET), Random Forest (RF) and K-Nearest Neighbor (KNN) are used to evaluate and predict the pH values. After the comprehensive evaluation, this study determined that the performance of the DT and RF model for pH prediction is high compared to the other algorithms in terms of accuracy. Further, the classifiers performance has improved by possessing feature scaling techniques like normalization and standardization. Results showed that the prediction accuracy of KNN and BNB with feature scaling outperforms the other algorithms.

Keywords: Agriculture, Soil Analysis, pH, Data mining, Classification.

(Received January 1st, 2021 / Accepted May 22nd, 2021)

1 Introduction

Agriculture is the foremost source of rural population of Tiruppur district. Red gravel and clay loamy soils are mostly found in Tiruppur district. Soils play a vital role in the agriculture ecosystem which provides us food, feed, fiber and fuel. The characteristics of each soil has the combination of these properties that is texture, structure, density, temperature, color, water holding capacity, consistency, and soil porosity. Soil that contains nutrients may initially be sufficient to sustain crop growth, but when they are continuously used for cultivation, nutrients are removed by crop harvesting. Nutrients removed by harvesting must be replaced for further crop production. Thus, maintaining the fertility of the soil is very important. Soil analysis is essential for farmers to apply nutrients in need to maintain soil fertility. Soil analysis is done by testing the soil samples to determine the soil characteristics, nutrients level, fertility, pH and soil contaminants. Among all, pH is one of the most important parameters in the soil analysis which determines the soil is acidic or alkaline in nature. As a result, predicting the type of pH is irresistible in determining the fertility of the soil, which increases yield and to avoid unnecessary chemical fertilizers. Data mining is a technology which can obtain the knowledge for agriculture development as well as predicting future trends of agriculture processes. It searches for hidden, valid, and probably useful and understandable patterns in large data sets. It is used almost in all areas where large amounts of data can be stored and processed. Data mining has no restrictions on the type of data to be analyzed [10]. In recent days, various researchers, data analysts and scientists have concentrated more on how mining techniques are used to analyze agricultural data using its techniques like classification, regression, clustering and other analysis methods [19]. The Data Mining task can be Predictive and Descriptive. Descriptive data mining focuses on providing the information based on past data in order to identify patterns. Descriptive data mining can be achieved using clustering, summarization, association rule, sequence discovery etc. On the other hand, predictive analytics can take both past and current data sets then it provides answers of the future values based on the known variables from both the data sets. Predictive analytics can be performed using classification, regression, time series analysis, etc. [2]. The main objective of this work is to determine the classification algorithm that yields the highest classification accuracy using soil pH level for the data set collected in 23 villages on Gudimangalam block of Tiruppur district, Tamilnadu. Further, this work attempts to improve the performance of the classifiers which ends with poor accuracy by utilizing the preprocessing techniques for feature scaling. This work examines the performance of the classification algorithms in the combination with preprocessing techniques in terms of performance evaluation measures.

Section 2 illustrates the different data mining techniques applied in the agriculture field. Section 3 explains the workflow used for predicting pH. Section 4 discusses feature scaling techniques. Section 5 explains the classification algorithms used in this work. Performance metrics for assessing model efficiency and results are discussed in Section 6 and 7. Section 8 concludes the research work and future enhancement.

2 Literature Review

Many studies have attempted to identify and solve agriculture problems using data mining techniques. In [15], Jeihouni et al., have used SAVI as an important covariate to predict soil moisture retention properties. Further the data mining techniques such as MARS, and GEP are used for digital soil mapping. Additionally, accuracy and performance of the models is evaluated through ME, MAE, RMSE, R2, and relative RMSE. Aarthi & Sivakumar [1] has developed a texture triangle classification system to plot texture classes automatically and it is combined with FCM to classify soil texture as sand, silt, clay and loam. In [9], soil CO2 emission based classification was carried out using data mining techniques and the performance of the different models were tested using the ROC curve. Jethva et al., [16] examined data mining techniques to analyze the soil dataset for fertilizer recommendation. At the end, an artificial neural network is suggested as the best classification algorithm. Similarly, a decision tree is best to analyze soil fertility. Samundeeswari et al., [23] has discussed the significant role of data mining techniques in agriculture and soil containments. Various classification algorithms such as Naive Bayes, J48 (C4.5) and JRip are utilized for soil classification. In [17], Majumdar et al., suggested PAM, CLARA, DBSCAN and Multiple Linear Regression to analyze the agriculture data set in Karnataka district for obtaining the finest parameters to maximize crop production. Transductive Support Vector Machine for predicting leaf disease has been used in [22]. In addition plant disease is detected through Latent Dirichlet Allocation and an Artificial Neural Network technique with the help of soil features and diseased plants. [6] has discussed different optimization techniques for crop planning. JRip and Naive Bayes are used in this work to predict soil type. Ramesh D et al., [21] used Multiple Linear Regression and Density-based clustering to predict crop yield. Chougule [8] proposed a Random Forest algorithm for crop recommendations based on type of crop along with region. Also, fertilizer recommendations based on N, P and K parameters from soil are developed using the K-Means algorithm. Crop damage due to grass grub insects is examined using Decision Tree, Random Forest, Neural Networks, Naive Bayes, Support Vector Machines and K-Nearest Neighbor algorithms [4]. Brenda et al., [7] has developed grape skin classification based on their chemical composition analyzed by ICP-MS using Multinomial Logistic Regression, K-Nearest Neighbor, Support Vector Machines, and Random Forest algorithms. Shivaranjani et al., [26] has reviewed supervised and unsupervised algorithms that help farmers to improve their yield by predicting weather and daily, monthly and yearly rainfall. Surya et al., [28] made a study on applying K-Means and K-Medoid clustering algorithms on agriculture dataset. In [3], an analysis on lime status level in the soil is done using different data mining techniques such as J48, Random Tree, JRip, OneR and Naive Bayes. Palepu et al., [18] provides a brief representation on data mining techniques for soil analysis and agriculture related problems. Hemageetha et al., [12] proposed Naive Bayes, J48, Bayesian Network and JRip to predict soil for crop cultivation based on the pH values.

3 Proposed methodology

Figure 1 represents the proposed workflow of the pH accuracy prediction. As a part of this work, 23 study sites were chosen for data collection. 17,185 soil samples data are used for this study. Feature scaling techniques like normalization and standardization are applied to achieve high quality data. The data is split using train_test_split for building the classification models. Training set is fitting for model construction and the testing set is for model validation. Hence, the data set collected is divided into the ratio of 70:30 for training and testing. Classification algorithms LR, BNB, DT, ET, RF and KNN were considered for the experimental study. Cross validation (CV) is done to access the predictive performance of the models. For this purpose, K Fold cross-validation is used; it divides all the samples into K numbers of sections, called folds. In this work, K is set as 10, i.e.10 Fold CV is performed. A confusion matrix is constructed from the predictive results to get the actual and predicted values obtained by the classification algorithms. These values are used to evaluate the performance measures such as accuracy, precision, recall and f1-score.

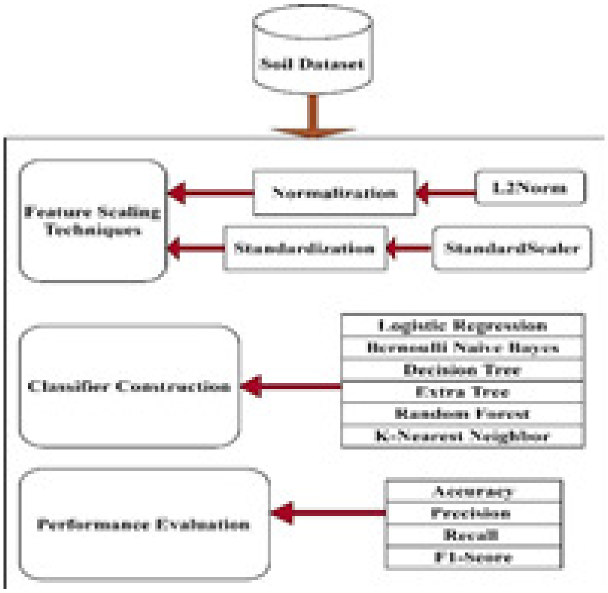

Figure 1: Proposed workflow

3.1 Soil Sample Selection

For assessing the quality of soil, a data set collected by the State Government of Tamilnadu (India) through the Department of Agriculture between the years 2015 & 2019 is used. The data set applied for analysis was collected from 23 villages on Gudimangalam block of Tiruppur district. These data sets are publically available in a soil health card portal which contains soil test values of corresponding parameters. Soil parameters considered in this study are Village Name, pH (soil reaction), EC (electrical conductivity), OC (organic carbon), P (phosphorus), K (potassium), Zn (zinc), Fe (iron), Cu (copper) and Mn (manganese). This work mainly focused on evaluating the performance of the classification algorithms over the soil data set of Gudimangalam block. pH status of the Tiruppur district is exhibited in figure 2.

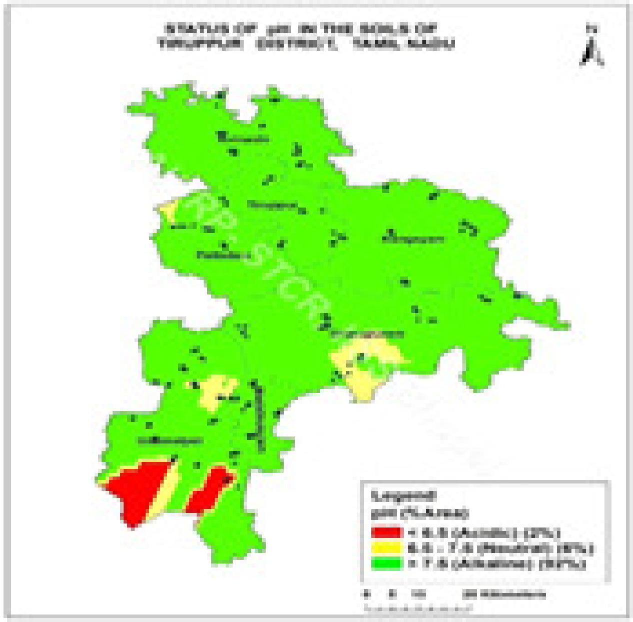
STATUS OF pH IN THE SOILS OF
TIRUPPUR DISTRICT, TAMIL NADU
N
Legend
pH (%Area)
🟥 < 6.5 (Acidic) (2%)
🟨 6.5 - 7.5 (Neutral) (8%)
🟩 > 7.5 (Alkaline) (90%)
0 5 10 20 Kilometers

Figure 2: pH levels in the soils of Tiruppur district

3.2 Soil pH Classification

pH has a great impact on soil biogeochemical processes. pH is considered as one of the most important soil parameters which affects both macro and micro nutrients. pH is measured in units and the scale goes from 0 to 14. The soil with higher pH or it with slightly too moderate alkaline, macro nutrients except phosphorus has increased and micro nutrients levels are reduced thereby it affects crop growth. Soils with low pH reduces macro and secondary nutrients whereas it increases Mn and Fe which are toxic to crops in excess. The soils of Tiruppur district are mostly nutrient deficient as they are alkaline in nature. Poor soil condition affects crop growth and limits yield. In order to plan the effective crop cultivation in an agricultural field, prior knowledge about the soil condition is mandatory. Hence, soil pH based classification is done using macro and micro nutrients [27]. The motivation towards this work is to diagnose the soil quality in Gudimangalam block of Tiruppur district. For this purpose, classification of pH has performed. pH levels for classification are given in Table 1. In this work only four pH levels such as Slightly Acidic (SA), Neutral (N), Slightly Alkaline (SAL) and Moderately Alkaline (MAL) are considered. Classification of soil pH is useful to determine the soil quality, available nutrients suitable crops and fertilizers.

Table 1: Soil pH & its values

Soil pH	*Ranges*
Extremely acidic	<4.4
Very strongly acidic	4.5-5
Strongly acidic	5.1-5.5
Moderately acidic	5-6-6
Slightly acidic	6.1-6.4
Neutral	6.5-7.5
Slightly alkaline	7.6-8
Moderately alkaline	8.1-9
Strongly alkaline	9.1-10
Very strongly alkaline	>10

3.3 Data Pre-Processing

Classification is the task to predict class labels of unknown samples. Each sample is described by a set of variables. Variables of the data set have to be easily interpreted by the algorithms. Data set with inappropriate values will result in false predictions. Therefore, data pre-processing becomes a vital and fundamental step to clean and transform the raw data into a suitable form for the data mining model. Feature scaling is performed during data pre-processing to adjust the data that has different scales to the same scale or range. This will reduce the dominance of one variable over others. If it is not done, a higher dominance variable is given a higher significance and a lower dominance variable is given to lower significance values irrespective of the unit of values i.e.,the reason to use feature scaling to bring all values in the same scale. Mostly, data mining algorithms perform well when numerical input variables are scaled to a standard range. Feature scaling methods used in this work are explained in section 4.

3.4 Classification

Predictive data mining/classification is the process to identify the likelihood of future outcomes based on historical data. The classification model is constructed using classification algorithms over training data set, once the model is constructed; it is applied over the test data set for validation [14]. In this work, the feature scaled data set is partitioned into training and test sets in the ratio of 70:30. Six classification algorithms such as LR, BNB, DT, ET, RF and KNN are taken for constructing classification models over training data set. In detail these six algorithms are described in section 5.

3.5 Performance Evaluation

Evaluating the model is the major part of building an effective data mining model. The predicted results are evaluated to compare the performances of the classification algorithms.10 fold cross validation is used to evaluate the classifiers performance. Confusion matrix is required to compute the accuracy of the data mining algorithms. Confusion matrix table which describes the test results of a prediction model. This contains both actual and predicted values. These values are used to calculate the accuracy score of the model. The performance evaluation metrics that are used to evaluate the classifiers are explained in section 6.

4 Feature Scaling

It is one of the significant pre-processing techniques used to standardize/normalize the input data. Since the data set considered in this work has attributes with different range values, it is quite important to perform normalization and standardization. They are the most popular feature scaling techniques used for scaling numerical values earlier to classification.

4.1 Normalization

Normalization is used to scale the values of a variable because when dealing with variables on a different scale leads to poor data models. So normalization is required to bring all the variables on the same scale in order to get better results.Generally, it is applied over the rows, not on the columns. L2 normalization or L2 norm, by default is applied on rows so that the values in a row have a unit norm.

- L2 norm

L2 norm (1) is defined as the normalization technique which is also known as least squares. It modifies the data set values in a way that in each row the sum of the squares will always be up to 1. It is better at minimizing the predicting errors. A value is normalized a

$$L2 - norm = \sqrt{x_i^2}, \forall i = 1...k \qquad (1)$$

4.2 Standardization

Standardization is a transformation that centers each input variable separately by subtracting the mean (µ) called centering and dividing by their standard deviation (σ) called scaling. It helps to improve the performance of the models.

- Standard Scaler

This standardizes a variable by subtracting the mean (µ) and dividing by standard deviation (σ) (2). It transforms the given distribution to a normal distribution having a zero mean and a standard deviation of one. All the variables will be of the same scale after applying the scaler.

$$Std - scalar = (x_i - \mu)/\sigma, \forall i = 1...n \qquad (2)$$

5 Classification Algorithms

The classification models are constructed over the soil samples using six classification algorithms to predict the soil quality of the Gudimangalam block. The models are built using Python which is an excellent tool for analyzing data with libraries such as Scikit-learn and StatsModels. Performance comparison between these models is analyzed using 10 fold cross validation over the test dataset. The performances of the predictive models are evaluated based on their obtained accuracy, precision, recall and f1-score values.

5.1 Logistic Regression (LR)

Logistic Regression is the one of the most common classification algorithms, used for analyzing binary and categorical target variables [13]. The connection between the categorical dependent variable and continuous independent variable is measured by changing the dependent variables into probability scores. Logistic regression is a traditional method used in many applications for building prediction models. Categorical variables could be binary or more than two levels called multinomial logistic regression (MLR). The example for logistic regression is shown in (3) as follows,

$$y = e^(b_0 + b_1 * x)/(1 + e^(b_0 + b_1 * x)) \qquad (3)$$

Where input value x is combined with coefficient values to predict the outcome value y. b0 is the bias.

5.2 Bernoulli Naive Bayes (BNB)

Bernoulli Naive Bayes is one among the family of Naive Bayes. The parameters used to predict the class variable takes up only two values because it assumes all the features as binary. This binary algorithm is used to build classifiers for multiclass problems. It splits the multiclass into binary class thereby creating one binary problem for each pair of classes. The decision rule for Bernoulli Naive Bayes is based on given Eq. (4),

$$P(x_i/y) = P(i/y)x_i + (1 - P(i/y)(1 - x_i) \qquad (4)$$

Where Bernoulli distribution has two outcomes(xi=1) or (xi=0). This algorithm gives more accurate and precise results.

5.3 Decision Tree (DT)

Decision tree is one of the predictive modeling algorithms used in data mining. This is a Tree like structure works on the principle of condition. Every node holds a variable, the link between the nodes holds the decision and every leaf node means the class label [25]. Decision trees are one among the popular classification algorithms to understand and interpret. During the decision making process, data with multiple attributes create difficulties to decide which attribute to place at the root node and this is a complicated step. For solving this attribute selection problem, Gini-index is one of the criteria that will calculate values for every attribute. These values are sorted and attributes are placed in an order. Formula for Gini-index (5) is given in following equation,

$$Gini - index = 1 - \Sigma_(i = 1)^n((p_i)^2)) \qquad (5)$$

5.4 Extremely Randomized Trees Classifier (Extra Tree Classifier)

Extra Trees Classifier (ET) is a type of ensemble learning technique. This method works by constructing random trees from the samples of the training data set. This random sample feature leads to the creation of multiple de-correlated decision trees. Predictions made by majority of votes obtained from the decision trees [11].

5.5 Random Forest (RF)

Random forest is a supervised learning algorithm that functions by constructing multiple decision trees that operate as an ensemble. Each branch of the tree represents a possible occurrence [14]. This algorithm will handle missing values, reduces the risk of overfitting and especially offers a high level of prediction accuracy [20]. Entropy (6) is a measure used to determine how nodes branch in a decision tree.

$$Entropy = -\Sigma_(i=1)^c(p_i * log_2(p_i) \tag{6}$$

Where pi is the probability of an element/class 'i' in data

5.6 K - Nearest Neighbor (KNN)

K - Nearest Neighbor is a most widely used algorithm especially for its simplicity. It uses simple measures to solve complex problems. It is a non-parametric approach that makes the algorithm more effective [24]. KNN algorithm checks how similar a data point is to its neighbor and classifies the data point into the class it is most similar to [5]. The KNN algorithm assigns a category to variables in the test dataset by comparing them to the variables in the training dataset. The parameter k is important in KNN algorithm because it decides how many neighbors will be chosen for KNN algorithm. In this study, K is assigned as 5. This means that the algorithm will consider five neighbors that are closest to the new data point. The closeness between the data points is calculated using Euclidean distance. Here, distance between (x1,x2...xn) and (y1,y2...yn) is measured using Euclidean distance measure (7) as follows.

$$Euclidean(x_i, y_i) = \sqrt{\Sigma_(i=1)^n((x_i - y_i)^2)} \tag{7}$$

Using (7), the distance between the points is calculated.

6 Performance Evaluation Metrics

Evaluation measures play a vital role in classification performance assessment. Choosing a right model is directed by the evaluation metric. Standard metrics such as classification accuracy or classification error are widely used for classification evaluation. Since the experiments are performed with six different algorithms, it is necessary to use performance evaluation metrics such as precision, recall and f1-score other than the accuracy.

- Accuracy

Accuracy (8) is the percentage of accuracy of the predictions made by the model.

$$Accuracy = No of correct predictions/Total no of predictions \tag{8}$$

- Precision

Precision (9) is the level up-to which the prediction made by the model is precise. It should ideally be 1 for a good classifier.

$$Precision = TP/TP + FP \tag{9}$$

TP means true positive and FP is denoted as false positive. FP is zero when precision is high.

- Recall

Recall (10) means the amount up-to which the model can predict the outcome.

$$Recall = TP/TP + FN \tag{10}$$

Both precision and recall to be one which means that the FP and FN (false negative) are zero.

- F1-Score

F1-Score (11) is the amount of data tested for the predictions. The higher the F1-Score, the most accurate the model is in doing predictions.

$$F1 - Score = 2TP/(2TP + FP + FN) \tag{11}$$

7 Experimental Results

The soil quality analysis of the experimental sites are estimated using six classifiers such as LR, BNB, DT, ET, RF and KNN. Accuracy, precision, recall and f1-score were used in order to evaluate the efficiency of the six models. The results of classifiers without feature scaling are shown in table 2.

Table 2: Performance of Classifiers without Feature Scaling

$Algorithms$	$Accuracy$	$Precision$	$Recall$	$F1 - Score$
LR	46	44	46	35
BNB	46	61	46	29
DT	100	100	100	100
ET	92	92	92	92
RF	100	100	100	100
KNN	57	58	57	57

From table 2, it is observed that the DT and RF achieves best performance in all measures. ET reached 92% in all measures. Following KNN, BNB performs better. LR has low accuracy among all models. According to all metrics, the DT and RF model is determined to be the best model in the soil pH prediction. Figure 3 shows the performance comparison of classifiers without feature scaling.

To improve the performance of classifiers other than DT and RF, the dataset is pre-processed using feature scaling techniques like normalization and standardization. Table 3 shows the results of the performance of classifiers using L2 normalized dataset. From the experiments, it is found that the ET model using normalized dataset achieves an accuracy of 72% which is slightly higher than other classifiers still it is 20% less than its original performance. Compared to the performances of the classifiers without feature scaling except ET, all other classifiers have shown improvement in their performances. With respect to feature scaling, models such as BNB and LR are showing very poor performance, which further needs improvement. Figure 4 shows the performance comparison of four models using the normalized data set.

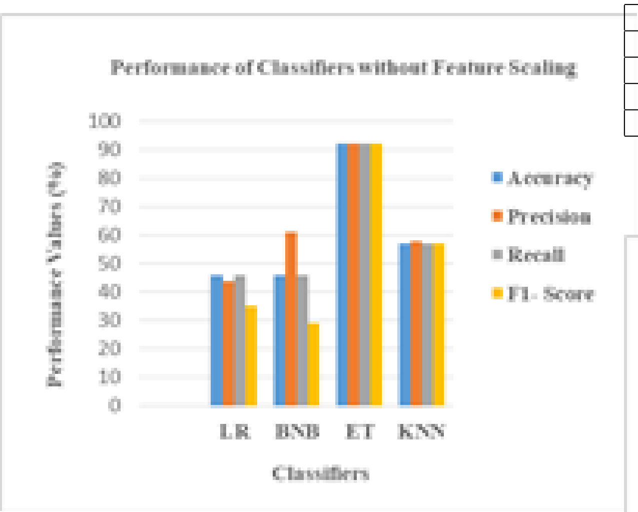

Figure 3: Performance Comparison of Classifiers without Feature Scaling

Table 4 shows the results of the performance of classifiers using standardized data set. From the results, it is found that BNB and KNN achieve good results in both terms of accuracy, precision followed by recall and f1-score. Figure 5 shows the comparison of four models after the normalization.

Among the four models, KNN model has the highest accuracy of 73%. Compared to normalization KNN performs well after standardization, i.e. the accuracy has increased nearly 14%. In addition, the evaluation accuracy of BNB is 72% and LR achieves 69% of accuracy which is much higher than its performance with normalization. ET has obtained an accuracy of 62% which is lesser than its original i.e without feature scaling. Overall KNN model has the highest soil quality assessment accuracy after applying standardization techniques. Figure 5 shows the comparison of classifiers.

Table 3: Performance of Classifiers using L2 Normalized dataset

$Algorithms$	$Accuracy$	$Precision$	$Recall$	$F1 - Score$
LR	47	45	47	37
BNB	46	61	46	29
ET	72	72	72	72
KNN	59	59	59	59

Table 4: Performance of Classifiers using Standardized Dataset

$Algorithms$	$Accuracy$	$Precision$	$Recall$	$F1 - Score$
LR	69	81	69	68
BNB	72	69	72	69
ET	62	70	62	63
KNN	73	83	73	74

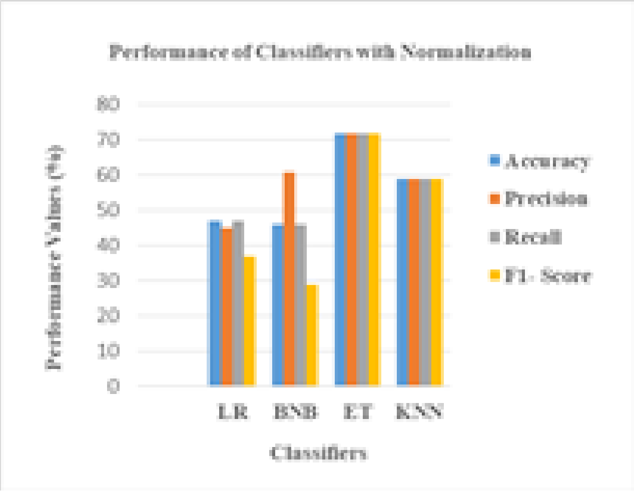

Figure 4: Performance Comparison of Classifiers using Normalized Dataset

Hence, from the experiments, it is proved that feature scaling techniques are helpful enough to improve

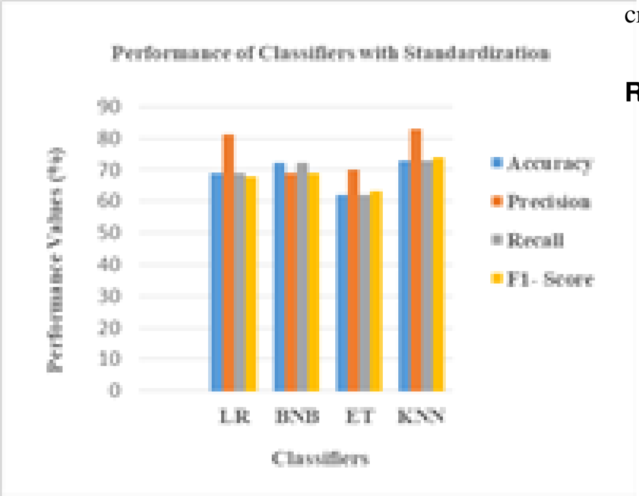

Figure 5: Performance Comparison of Classifiers using Standardized Dataset

the performance of the algorithms such as LR, BNB and KNN but not ET.

8 Conclusion and future work

Agriculture is largely dependent on soil. Soil quality is important and it is determined by many factors, all of which affect fertility. Soil loses its quality due to poor fertilization, irregular crop rotation, over-cultivation or soil contamination. A good soil will sustain the crop growth to a high standard and consistent yields in good quality, without the need for nutrient enhancements or chemical fertilizers. The proposed system uses the pH levels for assessing the soil quality using data mining algorithms. From the experiments, it is found that DT and RF achieve highest accuracy without feature scaling. Further to enhance the predictive accuracy of the classifiers, normalization and standardization are used for feature scaling. After normalization, ET achieves higher accuracy than other models but lesser than its original. When standardization is applied, the accuracy of KNN and BNB is improved than its original performance. From the experiments, it is also found that ET does not require any of the feature scaling techniques as its performance is deteriorated in feature scaled data set. The results of this study shows data mining models performed well and it is suited for soil quality analysis. This study assisted the Tamilnadu Government to take appropriate decisions to arrest dwindling quality of soil in Tiruppur and to improve crop production. In the future analysis, fertilizer recommendation systems are to be built to recommend appropriate fertilizers based on cropping patterns.

References

[1] Aarthi, R. and Sivakumar, D. An enhanced agricultural data mining technique for dynamic soil texture prediction. *Procedia Computer Science*, 171:2770–2778, april 2020.

[2] Agyapong, K., Hayfron-Acquah, J., and Asante, M. An overview of data mining models (descriptive and predictive. *International Journal of Software & Hardware Research in Engineering*, 4(5):53–60, 2016.

[3] Arunesh, K. and Rajeswari, V. Agricultural soil lime status analysis using data mining classification techniques. *Int. J. Adv. Technol. Eng. Sci*, 5(2):27–35, 2017.

[4] Ayub, U. and Moqurrab, S. A. Predicting crop diseases using data mining approaches: classification. *1st International Conference On Power, Energy And Smart Grid (Icpesg)*, pages 1–6, 2018.

[5] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. When is "nearest neighbor": meaningful? *International conference on database theory*, pages 217–235, 1999.

[6] Bhanudas, D. J. and Afreen, K. R. Prediction of soil accuracy using data mining techniques. *5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–5, 2019.

[7] Canizo, B. V., Escudero, L. B., Pellerano, R. G., and Wuilloud, R. G. Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. *Computers and Electronics in Agriculture*, 162:514–522, 2019.

[8] Chougule, A., Jha, V. K., and Mukhopadhyay, D. Crop suitability and fertilizers recommendation using data mining techniques. *Progress in Advanced Computing and Intelligent Engineering*, pages 205–213, 2019.

[9] Farhate, C. V. V., Souza, Z. M. d., Oliveira, S. R. d. M., Tavares, R. L. M., and Carvalho, J. L. N. Use of data mining techniques to classify soil co2 emission induced by crop management in sugarcane field. *Plos one*, 13(3):e0193537, 2018.

[10] Geetha, M. A survey on data mining techniques in agriculture. *International journal of innovative research in computer and communication engineering*, 3(2):887–892, 2015.

[11] Geurts, P., Ernst, D., and Wehenkel, L. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[12] Hemageetha and GM. Analysis of soil condition based on ph value using classification techniques. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 18(6):50–54, 2016.

[13] Hilbe, J. M. Logistic regression models. *CRC press*, 2009.

[14] Issad, H. A., Aoudjit, R., and Rodrigues, J. J. A comprehensive review of data mining techniques in smart agriculture. *Engineering in Agriculture, Environment and Food*, 12(4):511–525, 2019.

[15] Jeihouni, M., Alavipanah, S. K., Toomanian, A., and Jafarzadeh, A. A. Digital mapping of soil moisture retention properties using solely satellite-based data and data mining techniques. *Journal of Hydrology*, 585:124786, 2020.

[16] Jethva, J. M., Gondaliya, N., and Shah, V. A review on data mining techniques for fertilizer recommendation. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, IJSRCSEIT*, 3(1), 2018.

[17] Majumdar, J., Naraseeyappa, S., and Ankalaki, S. Analysis of agriculture data using data mining techniques: application of big data. *Journal of Big data*, 4(1):1–15, 2017.

[18] Palepu, R. B. and Muley, R. R. An analysis of agricultural soils by using data mining techniques. *Int. J. Eng. Sci. Comput*, 7(10), 2017.

[19] Patel, H. and Patel, D. A brief survey of data mining techniques applied to agricultural data. *International Journal of Computer Applications*, 95(9), 2014.

[20] Qi, Y. Random forest for bioinformatics. *Ensemble machine learning*, pages 307–323, 2012.

[21] Ramesh, D. and Vardhan, B. V. Analysis of crop yield prediction using data mining techniques. *International Journal of research in engineering and technology*, 4(1):47–473, 2015.

[22] Sabareeswaran, D. and Sundari, R. G. A hybrid of plant leaf disease and soil moisture prediction in agriculture using data mining techniques. *Int. J. Appl. Eng. Res*, 12(18):7169–7175, 2017.

[23] Samundeeswari, K. and Srinivasan.K. Data mining techniques in agriculture prediction of soil fertility. *International Journal of Scientific Engineering Research*, 8(4):2229–5518, 2017.

[24] Saranya, N. and Mythili, A. Classification of soil and crop suggestion using machine learning techniques. *International Journal of Engineering Research Technology (IJERT)*, 2020.

[25] Sharma, H. and Kumar, S. A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, 5(4):2094–2097, 2016.

[26] Shivaranjani, M. and Karthikeyan, K. A review of weather forecasting using data mining techniques. *International Journal of Engineering and Computer Science*, 5(12):19784–19788, 2016.

[27] Sirsat, M., Cernadas, E., Fernández-Delgado, M., and Khan, R. Classification of agricultural soil parameters in india. *Computers and electronics in agriculture*, 135:269–279, 2017.

[28] Surya, P. and Laurence Aroquiaraj, I. Performance analysis of k-means and k-medoid clustering algorithms using agriculture dataset. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 6(1), 2019.