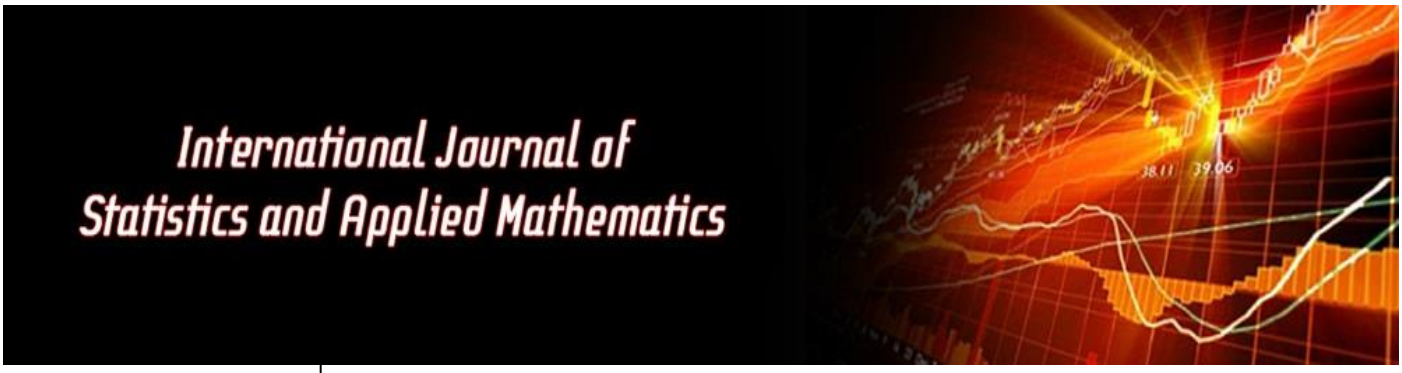
International Journal of Statistics and Applied Mathematics 2023; SP-8(5): 475-481



ISSN: 2456-1452
Maths 2023; SP-8(5): 475-481

https://www.mathsjournal.com
Received: 10-06-2023
Accepted: 15-07-2023

Baby Akula
College of Agriculture, PJTSAU, Hyderabad, Telangana, India

Dr. K Indudhar Reddy
AICRP, WIA, PJTSAU, Rajendranagar, Hyderabad, Telangana, India.

Divya N
Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.

Parmar RS
COAIT, Anand Agricultural University, Anand, Gujarat, India.

Corresponding Author:
Dr. K Indudhar Reddy
AICRP, WIA, PJTSAU, Rajendranagar, Hyderabad, Telangana, India.

Advances in soil fertility classification: Data mining approach

Baby Akula, Dr. K Indudhar Reddy, Divya N and Parmar RS

DOI: https://doi.org/10.22271/maths.2023.v8.i5Sg.1240

Abstract
Agriculture is as old as civilization. Indian agriculture heritage witnessed nomadic shifting cultivation to present precision agriculture feeding its ever-burgeoning population. Indeed a magical feat that no other country in the world can partake in India in terms of its agriculture production.
Soil fertility, an important dimension of soil productivity braced up India from starvation to self-sufficiency. This paper focuses on soil fertility classification of soil dataset of Ranga Reddy district of Telangana using data mining techniques *viz*., Tree based – Random Forest, Random Tree and; Lazy based -IBK and K Star; and Bayesian-based Naïve Bayes. The soil fertility dataset extended to 2,408 instances comprising 12 attributes of soil parameters which included soil physicochemical properties, and macro, secondary and micronutrients of soil collected from selected model villages of the Ranga Reddy district. Soil the performance of each model was examined in terms of correctly classified instances, incorrectly classified instances, Receiver Operating Characteristic (ROC) Area, Kappa statistic., mean absolute error, Root mean squared error, Relative absolute error and Root relative squared error. The classification algorithm, the Random Forest model had achieved the highest prediction accuracy of 93.69%, sensitivity of 0.937 and precision of 0.902 and F1 score of as compared with the rest of the models.

Keywords: Data mining, Indian agriculture heritage, soil fertility, ROC

Introduction
Fundamentally, in a dense populous country like India, any effort that promotes agricultural development assumes significance just for the reason that no other country in the world can feed the mouths of India. Agriculture is as old as civilization in India. Since ages, different shades of cultivation practices including nomadic shifting cultivation to advanced precision agriculture were being practiced to meet huge food demand. During the entire saga of agriculture, soil fertility was the driving factor that enhanced food production disproving the Malthusian theory that India would starve to death. Estimation of soil fertility is back back-breaking and time-consuming process and thus data mining approach of soil fertility classification on which the present paper focuses, gains importance in better decision making about fertilization, irrigation and other crop management practices. In the future, data mining or machine learning techniques are
Likely to continue as a key driver of agriculture development in India and elsewhere of the world.
Data Mining is the process of analyzing, extracting and predicting meaningful information from enormous data to various patterns. Rajeswari and Arunesh (2016) [8] made a comparative analysis of three algorithms namely Naïve Bayes, JRip and J48. JRip correctly classified the maximum number of instances as compared with other two models. Chiranjeevi and Ranjana (2018) [2] conducted a comparative analysis of two algorithms namely Naive Bayes and J48. However, Raunak (2018) [9] recommended Naive Bayes to predict crop growing in particular soil types based on their studies. Kalekar *et al.* (2018) [5] suggested J48 decision tree algorithm as it showed an accuracy of 87.5% in classifying the soil fertility to use for fertilizer recommendation.

Jethva *et al.* (2018) [4] made a comparative analysis of algorithms like J48, Naïve Bayes, JRIP, and K-Means classifier algorithms and concluded that the decision tree algorithm performed the best to analyze soil fertility conditions. Prajapati *et al.* (2019) [7] proposed three approaches of nonparametric classifiers; fast K-nn (K-nearest neighbors), training set reduction techniques and hybrid approach. It was observed that K-nn classification technique was better than other approaches.

Methodology

In this research, soil fertility datasets were collected from the Department of Agriculture, Ranga Reddy District, and Telangana. The soil fertility parameters were collected from twenty-two model villages from each mandal of Ranga Reddy district which has geographic importance by virtue of its proximity to Hyderabad, the capital city of Telangana state, a newly formed 29th state of India in 2014. Intensive peri-urban agriculture is way of life to the majority of farmer's hence high demand for fertilizers and other inputs. Farmers are tempted for misuse of fertilizers for higher productivity because of easy marketability and roadworthiness. The dataset consisted of soil fertility data having 13 attributes namely pH value of soil (pH), Electric conductivity (EC), Organic Carbon (OC), Nitrogen (N), Phosphorous (P_2O_5), Potassium (K_2O), Sulphur(S), Zinc(Zn), Iron(Fe), Copper(Cu), Manganese (Mn), Boron(B), Fertility Index (FI). And thus total soil fertility dataset had 2,804 instances. In Fig. 1, the structural design of soil fertility classification system is being given.

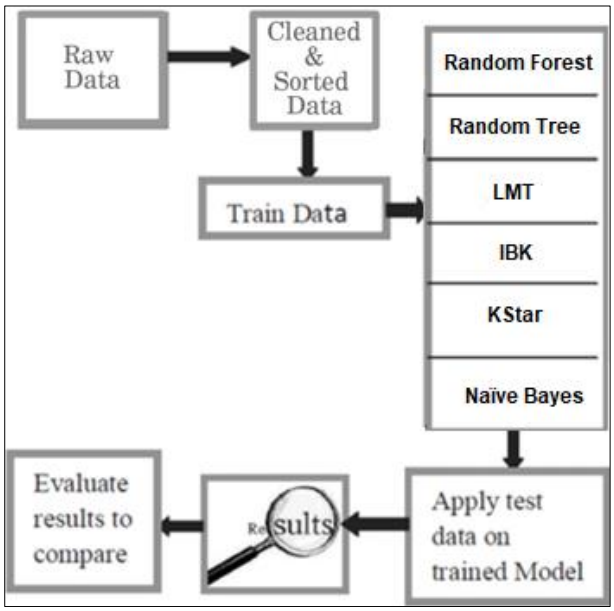


Fig 1: Structural Design of Soil Fertility Prediction System

Steps followed while classification (prediction) soil fertility

1. This dataset was created in an Excel sheet with.CSV extension.
2. The raw data were cleaned and sorted.
3. The fertility index was class label which was categorized as L - Low, VL- Very Low, M - Medium, H - High, VH - Very High, D - Deficient, S - Sufficient, AS - Acid sulfate, SrAc- Strongly acidic, HAc - Highly Acidic, MAc - Moderately Acidic, SlAc - Slightly Acidic, N - Neutral, MAl - Moderately Alkaline, SlAl - Strongly Alkaline.
4. Open-source Data mining tool WEKA version 3.8.1 was used for the classification of soil fertility.
5. Min-Max Normalization technique was used to normalize the soil fertility dataset which reduces large variation of estimation (eq.1). Normalization is the process of changing the values of data,a where to find new range from an existing range.

Min-Max normalization technique:

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \quad (1)$$

Where v is the respective value of the attribute
V^1 is Min-Max Normalized data one
$\min_A$ is the respective Minimum of value of the attribute
$\max_A$ is the respective Maximum value of the attribute

6. As a next step, in data mining, feature selection was followed. It is also known as predictor selection. Feature selection techniques were used for the simplification of models to make them easier to interpret by researchers, shorter training times, to avoid the curse of dimensionality and to reduce over fitting. Thus, 12 factors were considered as the predictors and soil fertility index was taken as the target variable. As few predictors may be superfluous, affecting estimation of soil fertility index. So, from 12 predictors; only six (6) predictors namely pH value of soil (pH), Electric conductivity (EC), Organic Carbon (OC), Nitrogen (N), Phosphorous (P_2O_5), Potassium (K_2O) had positive and strong association with target variable were selected using feature selection algorithm namely "cfsSubsetEval".
7. Cross-validation procedure was used to assess machine learning models on a limited data sample. The k-fold cross-validation has a parameter called k that refers to the number of groups that a given data sample is to be split into. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming a 10-fold cross-validation. Cross-validation is mostly used in applied machine learning to estimate the skill of a machine learning model on unseen

data. That is, to use some degree of sample in order to estimate how the model is projected to perform in general when used to make predictions on data not used during the training of the model.

8. Classification is a method where we classify data into a given number of classes. A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data. The different classification techniques *viz*., Random Forest, Random Tree, LMT, IBK, K Star and Naïve Bayes were then implemented over the trained data (Table 1).

Table 1: List of Classification Algorithms in Weak 3.8

Category	Algorithm	Classification Principle
Tree Based	Random Forest	Collection of decision trees focused on random features extracted from bootstrapped data
	Random Tree	It is an ensemble learning algorithm that generates lots of individual learners. It employs a bagging idea to construct a random set of data for constructing a decision tree
	LMT	Classification trees with the method of logistic regression on the leaves.
Lazy Based	IBK	Learner-based instance utilizes the class of the closest k instances.
	K star	The learner focused on an instance utilizing an entropic distance scale.
Bayesian Based	Naïve Bayes	Conventional Bayesian probabilistic classification

9. Confusion Matrix was calculated. A confusion matrix is a technique for summarizing the performance of a classification algorithm (Fig.2).

- **TP:** Number of instances where the system detects for a condition when it is present.
- **TN:** Number of instances where system does not detect a condition when it is absent.
- **FN:** Number of instances where the system does not detect a condition when it is present.
- **FP:** Number of instances where the system detects a condition when it is really absent.

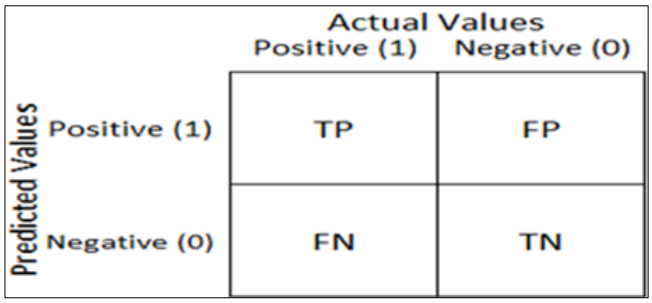


Fig 2: Confusion Matrix

10. Correctly Classified Instances, Incorrectly Classified Instances, Receiver Operating Characteristic (ROC) Area, Precision-Recall Curve (PRC) Area, Kappa Statistic, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error and Root Relative Squared Error, True Positive Rate, False Positive Rate, Precision, F-Measure and MCC values were taken into consideration for each case.

- **RMSE:** It is defined as the difference between the values predicted by the model and the actual values noted
- **MAE:** It is another factor in statistics that measures the difference between two continuous variables.
- **RAE:** This measure gives the total absolute error between the variables
- **Accuracy:** It is defined as the overall success rate of the classifier (TP+TN) / (TP+FN+FP+TN).
- **Sensitivity:** It is defined as a percentage of correctly classified instances. It is True Positive Rate (TPR = TP/ (TP/FN)).
- **Specificity:** It is defined as a percentage of incorrectly classified instances. It is a True Negative Rate (TNR = TN/ (FP+TN)).
- **ROC Area:** It is a performance measurement for classification problems at various threshold settings.
- **F1 score:** Test's accuracy
- **MCC:** Measure the quality of classification

11. Thereafter performance was measured using three factors namely Sensitivity, Specificity, and Accuracy.
12. The results of each algorithm were noted from WEKA and compared with each other.

Results

The results presented in Table 2 indicated that the Tree-based and Lazy based models have better performance than Bayesian based model. In tree-based models among the three algorithms examined, RandomForest performed better as compared with RandomTree and LMT in terms of more numSSber of correctly classified instances and ROC. The kappa statistic was nearer to one and relatively less error percentage (MAE, RMSE, RAE) by RandomForest indicating its better performance as compared with than other fitted algorithms. In Lazy based models,among two algorithms that were studied namely IBK and KStar, IBK was found to be better over KStar as the former recorded less error in predicting soil fertility. The Bayesian based Naïve Bayes performance was least as compared with the rest of all the algorithms tested.

Table 2: Performance of Different Classification Model Algorithms

Performance Error	Tree Based			Lazy Based		Bayesian Based
	Random Forest	Random Tree	LMT	IBK	K Star	Naïve Bayes
Correctly Classified Instances (Prediction Accuracy, %)	93.70	93.69	93.54	93.69	88.59	70.36
Incorrectly Classified Instances (%)	6.30	6.31	6.46	6.31	11.41	29.64
Receiver Operating Characteristic (ROC)	0.969	0.967	0.951	0.967	0.940	0.938
Kappa Statistic	0.9365	0.9364	0.9349	0.9364	0.8848	0.7005
Mean Absolute Error (MAE)	0.0001	0.0003	0.0009	0.0002	0.0024	0.0025
Root Mean Squared Error (RMSE)	0.013	0.015	0.017	0.014	0.029	0.030
Relative Absolute Error (RAE, %)	5.389	6.457	22.178	5.905	56.975	49.493
Root Relative Squared Error (%)	29.537	32.512	36.396	29.688	62.418	64.112

The Fig. 3 shows the prediction accuracy of different classification models. Out of six models used in this research work, Random Forest showed maximum soil fertility predictability of 93.70% over the other classification models. It was followed by Random Tree and IBK with 93.69% and 93.54% respectively. LMT followed by K Star exhibited lower accuracy of 93.69% and 88.59%, respectively, while, Naïve Bayes classification recorded lowest soil fertility predictability of 70.36%.These results can be corroborated with the findings of Keerthan *et al.* (2020) [6] And Elhamayed (2016) [3] who also stated that Random forest type classification algorithm scored highest accuracy in grading the soil, b based on its nutrient criteria.

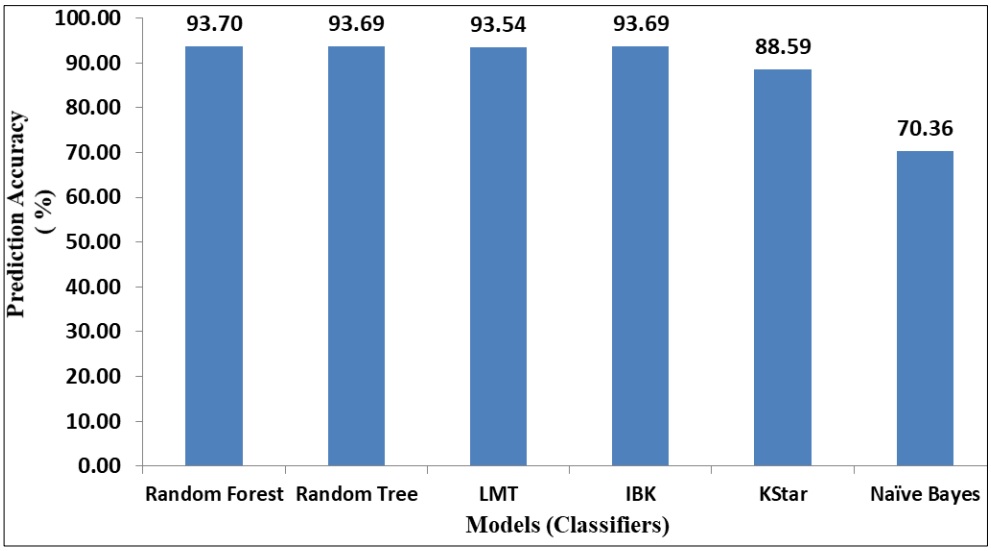


Fig 3: Prediction Accuracy of Different Classification Models

Fig. 4 depicts the error result of the different classification models. Random Forest had the minimal mean absolute error (MAE) of 0.0001 and root mean squared error (RMSE) of 0.013. During the prediction processes. In contrast, the Naïve Bayes classification had the highest error rate with 0.0025 and 0.030 of MAE and RMSE respectively as compared with prediction algorithms.

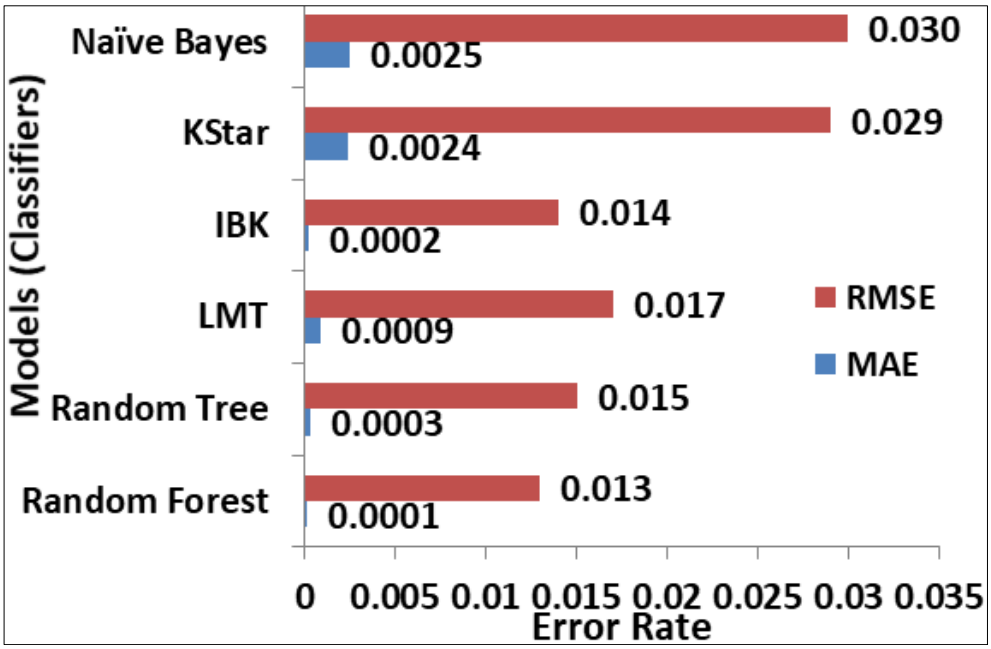


Fig 4: Error Results of Classification Models

Fig. 5 explains the true positive rate of different classification models. Out of six models used in this research work, Random Forest had a better true positive rate as compared with other classification models with 0.937, followed by Random Tree and IBK with 0.936, LMT with 0.935, and Kstar with 0.886. Naïve Bayes classification had the lowesttrue positive rate with 0.704.

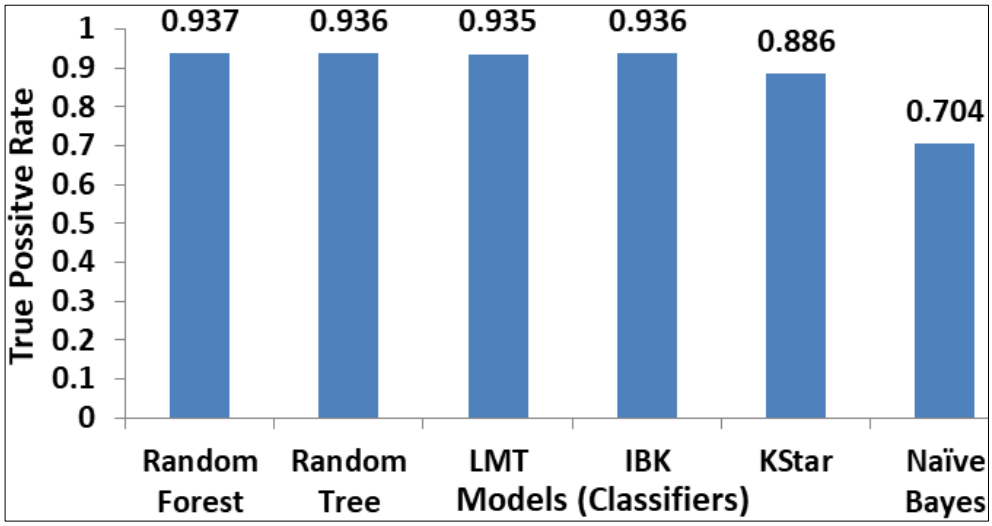


Fig 5: True Positive Rate (Sensitivity) of Classification Models

The Fig. 6 shows the false positive rate of different classification models. Out of six models used in this research work, Random Forest, Random Tree, and IBK showed the lowest false positive rate with 0.001 followed by LMT with 0.002 and Kstar with 0.003. Naïve Bayes classification had the highest false positive rate of 0.004.

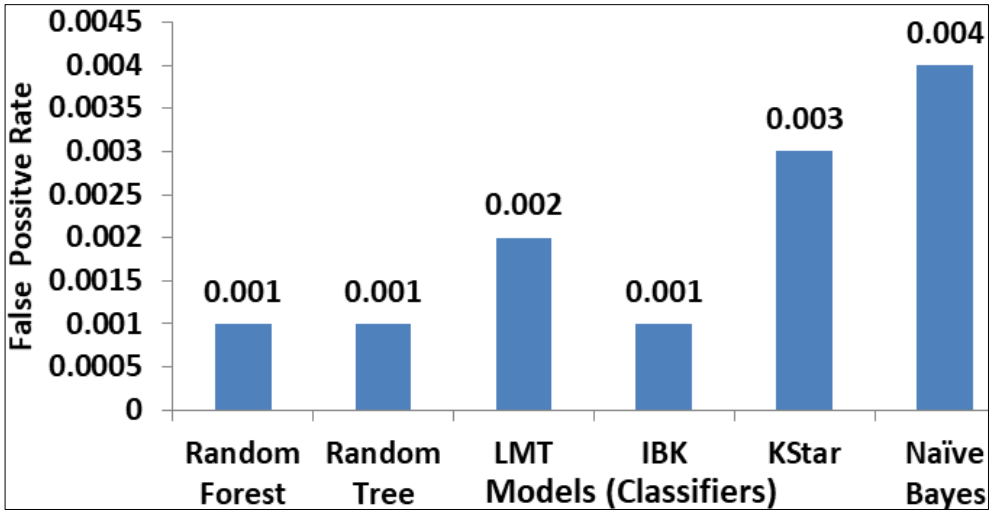


Fig 6: False Positive Rate (Specificity) of Classification Models

The Fig. 7 depicts the kappa statistics different classification models. Random Forest, Random Tree, IBK and LMT exhibited nearer to one value of kappa statistics. Naïve Bayes classification has the lowest kappa statistics with 0.7005.

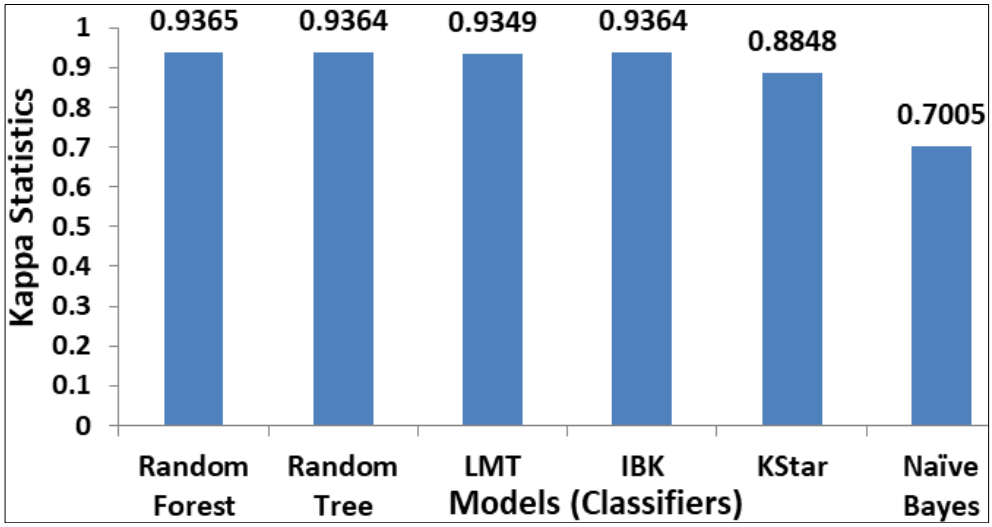


Fig 7: Kappa Statistics of Classification Models

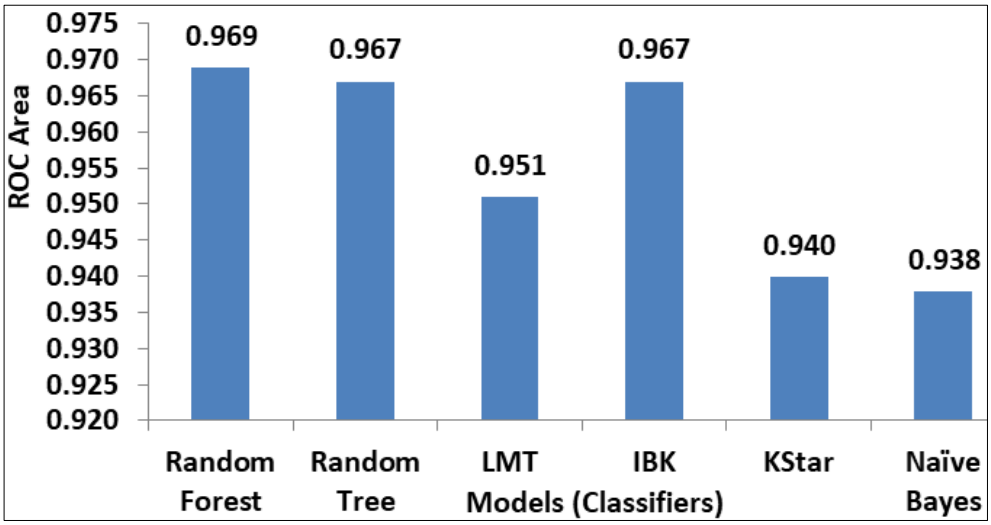


Fig 8: ROC Area of Classification Models

Table 3: Comparison of the statistics of the Classification Models

Parameters	Classifier Model					
	Tree Based			Lazy Based		Bayesian Based
	Random Forest	Random Tree	LMT	IBK	K Star	Naïve Bayes
Accuracy	93.69%	93.69%	93.54%	93.69%	88.59%	70.36%
Sensitivity	0.937	0.937	0.935	0.937	0.886	0.704
Precision	0.902	0.899	0.800	0.887	0.799	0.521
Specificity	0.001	0.001	0.002	0.001	0.003	0.004
F1 score	0.917	0.916	0.850	0.910	0.837	0.594
MCC	0.918	0.917	0.849	0.911	0.839	0.602

The Fig. 8 explains the ROC Area for different classification models. The trend is similar to other predictability parameters. Random Forest recorded the highest ROC Area with 0.969, followed by Random Tree and IBK with 0.967, followed by LMT with 0.951, and followed by K star with 0.940. Naïve Bayes had has the lowest area with 0.938.

The Table 3 explains the statistics after applying the six classification models on the soil fertility dataset. The fitted Random Forest model had achieved the highest prediction accuracy of 93.69%, the sensitivity of 0.937 and a precision of 0.902. On the other hand, Naïve Bayes had achieved the lowest prediction accuracy of 70.36%, sensitivity of 0.704 and precision of 0.521. In case of specificity, the Random Forest model obtained the lowest specificity of 0.001 and Naïve Bayes model had obtained the highest specificity of 0.004. F1 score and Mathews Correlation Coefficient were computed to measure the test's accuracy and quality of classification respectively. The Random Forest had achieved the highest F1 score of 0.917 and MCC of.0.91 while, Naïve Achieved the lowest F1 score of 0.594 and MCC of 0.602. Random Tree > IBK > LMT>K stat showed descending order of F1 score and MCC values. Soil fertility categorization based on different classification models indicated that J48 and Random Forest classifier models were found to be more effective in the classification of soil fertility rate by Bhuyar *et al.* (2014) [1] in the Aurangabad region, India.

Conclusion

Model predictive approach of agriculture is a reality now with strides in information technology and data mining is such a scientific tool, which provided essential knowledge from the database such as soil fertility data of Range Reddy district of Telangana. Based on all the benchmarks used to measure the models employed in this study, it was discovered that Tree based Random Forest model with 93.69% accuracy, 0.937 sensitivity, 0.902 precision, 0.001 specificity, 0917 F1 score and 0.918 MCC is the most appropriate in terms of classification based on this data. Thus, it was observed that Random Forest is best fitted model to classify the soil fertility dataset.

Future Work

Soil fertility prediction can be much more simplified by developing a GUI toolbox based on the Random Forest algorithm.

References

1. Bhuyar V. Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District. International Journal of Emerging Trends & Technology in Computer Science. 2014;3(2):200-203.
2. Chiranjeevi MN, Ranjana BN. Analysis of Soil Nutrients using Data Mining Techniques. International Journal of Recent Trends in Engineering & Research. 2018;4(7):103-107.
3. Elhamayed SA. Enhancement of Agriculture Classification by Using Different Classification Systems. International Journal of Computer Applications. 2016;3(1):08-12.
4. Jethva JM, Gondaliya N, Shah V. A Review on Data Mining Techniques for Fertilizer Recommendation. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2018;3(1):2456-3307.
5. Kalekar A, Vispute S, Kokane P, Kamble M, Bokefode K. Automated Generation and Analysis of Soil Health Card and Calculation of the Village Soil Fertility Index. International Journal of New Technologies in Science and Engineering. 2018;5(3):118-125.

6. Keerthan Kumar TG, Shubha C, Sushma A. Random Forest algorithm for soil fertility prediction and grading using Machine learning. International Journal of Innovative Technology and Exploring Engineering. 2020;9(1):1301-1303.
7. Prajapati BP, Kathiriya DR. A Hybrid Machine Learning Technique for Fusing Fast k-NN and Training Set Reduction: Combining Both Improves the Effectiveness of Classification. In Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore. 2019;(174):229-240.
8. Rajeswari V, Arunesh K. Analysing Soil Data using Data Mining Classification Techniques. Indian Journal of Science and Technology. 2016;9(19):1-4.
9. Raunak J. Applying Naive Bayes Classification Technique for Classification of Improved Agricultural Land Soils. International Journal for Research in Applied Science & Engineering Technology. 2018;6(5):189-193.