International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Prediction of Best Suitable Crop using Machine Learning Technique

Vaishali Kadwey, Anil Kumar Gupta

Submitted: 07/10/2024 **Revised:** 20/11/2024 **Accepted:** 29/11/2024

Abstract: The selection of best crop for cultivation, suitable according to agronomic and environmental factors at the particular area is a critical and responsible decision for the farmers. As agriculture in our country play very significant role in nation building by facilitating food export and providing major employment. The use of machine learning technology in agriculture field became boon because its nemours applications such as crop disease and weed detection, smart irrigation system, crop monitoring system and crop yield prediction etc. the study is focused on development of the ensemble regression model by leveraging the robust power of machine learning. Random Forest, Gradient Boosting and Linear Regression are used as base models. The Stacking and Voting techniques are used for development of proposed hybrid Models HVM1and HSM2, the performance matrix RMSE, MSE, MAE and R2 Score evaluated for the HVM1 and HSM2 has been compared. It is observed that the hybrid model HSM2 has highest R2 Score and lowest RMSE, MSE, MAE. The proposed hybrid model HSM2 helps farmer for selecting most suitable crop for cultivation for specific soil nutrients and environmental conditions,

Keywords: *Machine Learning, Linear Regression, Random Forest Regression, Gradient Boosting Regression, Voting, Stacking.*

1. Introduction

In nation-building the Agriculture plays a significant role. It influences a nation's economic, by increasing productivity, generating employment opportunities for a significant portion of the population in various sector. The developments in agriculture directly contribute to national economic growth and GDP. Advances and innovations in agricultural technology enhance food production and propel a nation into a more competitive global market. The India is the one of the largest agricultural producer country in the world. In India the cultivation of a wide variety of crops became possible because of diverse soil, land types and climates. For ensuring food security and optimizing agricultural practices across the country, proper understanding of environmental and agronomic factors for the cultivation and proper growth of crops is essential. Machine learning brings revolution in agriculture by adopting smarter farming practices. Machine learning (ML) is transforming agriculture by enhancing productivity, improves decision-making at every stage from cultivation to harvesting, and reduces resource usage, making agriculture more eco-friendly and sustainable. There are number of applications of Machine learning in agriculture few of them are Field Monitoring, Yield Prediction, Weather Forecasting, Crop Disease, Weed and Pest Detection and management, Soil Health Monitoring. The implementation of machine learning applications in real world helps farmers, policy makers, government and agro based businesses, and it is responsible for sustainable development in the field of agriculture.

2. Literature Review

[1] this research presents predictive model by Integrating the power of machine Learning, optimization, and agronomic insight, the model has three features in first feature it predict yield of Maize and Soybean for three Midwest states (Iowa, Indiana, and Illinois) of US and achieve 8% less root mean square error. Second feature it identified about environment by management interactions for corn and soybean yield. in third it analyze the crop yield on the basis of impact of soil, weather and management. [2] the research carried out by N Bali studied near about 100 papers related to the crop and technology used it is observed that machine learning role in impact agronomic and environmental factors on crop 17 papers, 43 papers are found about machine role in for crop yield prediction. 15 papers are used deep learning

Department of Computer Science & Applications
Barkatullah University Bhopal
vaishali2498@gmail.com
Department of Computer Science & Applications
Barkatullah University Bhopal
akgupta_bu@yahoo.co.in

technique for the study in the field. The main aim of study has to find out the various techniques used and factors affecting on crop yield. [3] To increase the crop yield and minimized the loss from pest and weed. The Markov random field (MRF) is used. The data collected from Australian farms.

[4] They investigated features, analyzed algorithms and retrieve seven features from various databases. The Regressor algorithm like Decision Tree, Ad boost, neural network, Gradient Boosting. Random Forest, Bagging classifier are used for training mode. The Ensemble technique using decision Tree and AdaBoost performs excellent in crop prediction for given location and environmental conditions [5] The study carried out by R Medar has implement crop selection method for getting maximum yield using WEKA tools. The framer gets help in selecting best crop for cultivation according to season and type of land. The K-Nearest neighbor and Naïve Bayes methods have been used and Naïve Bayes model performed best.

[6] designed the system for betterment of farmers using machine learning, the model suggest most suitable crop particular agronomic and climatic condition. The system gives information about the requirement of seeds for cultivation and quantity of fertilizer. SVM is used for prediction of rainfall and Decision Tree is used for crop prediction.

[7] In the studies, three countries (the Netherlands (NL), Germany (DE) and France (FR)) and five crops (soft wheat, spring barley, sunflower, sugar beet and potatoes) are considered for yield prediction. in the research machine learning combined with agronomic principles of crop modeling to develop machine learning baseline for crop yield forecasting on large scale. It is observed that machine learning baseline performed excellent as compare to MCYFS (MARS Crop Yield Forecasting System). [8] Used 22 ha field in Bedfordshire, UK for predicting wheat yield. The data collected from satellite imagery crop growth characteristics and on-line multi-layer soil data. XY-Fs (XY-fused Networks), CP-ANNs (counter-propagation artificial neural networks) and SKNs (Supervised Kohonen Networks) models are used for predicting wheat yield. the SKN model perform excellent with 81.65% accuracy, the CP-ANN and XY-F has 78.3% and 80.92% showing accuracy respectively.

[9] The major objectives of the study have to developed machine learning-based palm oil yield prediction model. The machine learning RF, LR, NN and some Deep learning model such as DNN, LSTM and CNN are used for palm crop yield prediction.

[10] The author suggested IoT based smart forming system called WPART for predicting crop productivity and drought. The WPART performs excellent for crop productivity and drought classification with maximum accuracy than existing standard algorithms. [11] This study comprises about 39 papers over various areas of crop management using machine learning like disease detection, yield prediction, soil and water management and weed detection. The study proposed model designed using SVM and digital images processing to determine maturity stages of a crop wheat crop.

3. Methodology

The dataset for proposed study has been taken from secondary source kaggle and government site https://www.soilhealth.dac.gov.in the has information about 22 different crops, soil nutrients like Nitrogen (N), Potassium (K), Phosphors (P), soil pH, environmental factor like Humidity, Temperature, Rainfall.

3.1 Data Preprocessing:

The dataset consider for the study has soil nutrients like N, P, K, pH, and environmental factor like Humidity, Temperature, Rainfall; all this features are numerical and scaled by MinmaxScaler and standardized by StandardScalar. Dataset also has one categorical feature label (name of crops) has been preprocessed by LabelEncoder of scikit learn.

3.3 Model Development

The processed data is split into subset as training and testing dataset in 80:20 ratios through train test split function of sklearn. The 80 percent data is used to train models developed by Linear regression, Random Forest, Gradient Boosting and Hybrid Models HVM1 (Hybrid Voting Model1) and HSM2 (Hybrid Stacking Model2) developed using ensemble Stacking and Voting techniques.

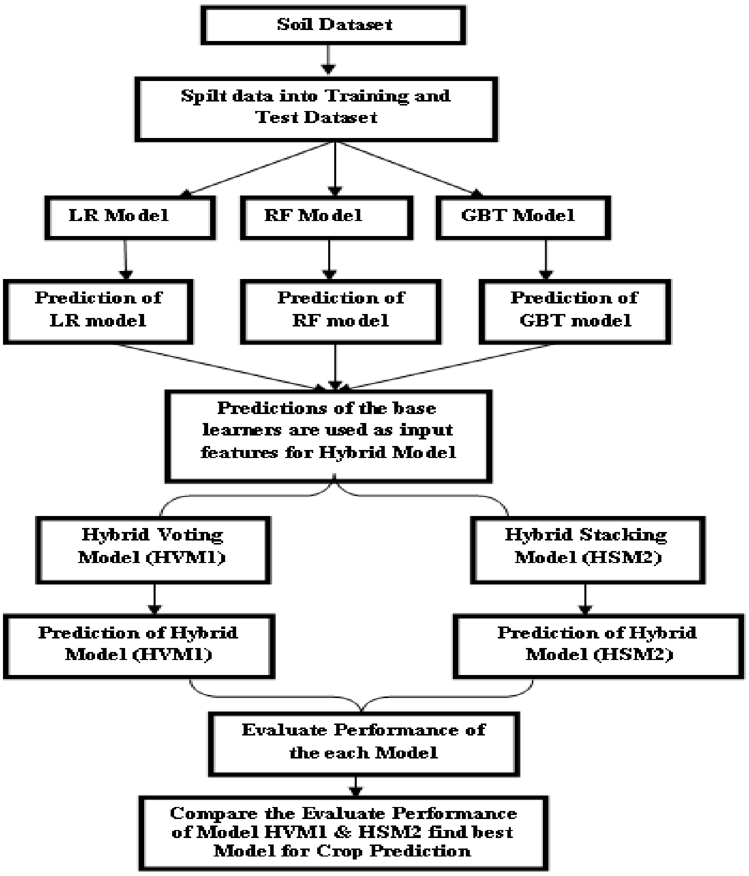
Soil Dataset
Spilt data into Training and Test Dataset
LR Model
RF Model
GBT Model
Prediction of LR model
Prediction of RF model
Prediction of GBT model
Predictions of the base learners are used as input features for Hybrid Model
Hybrid Voting Model (HVM1)
Hybrid Stacking Model (HSM2)
Prediction of Hybrid Model (HVM1)
Prediction of Hybrid Model (HSM2)
Evaluate Performance of the each Model
Compare the Evaluate Performance of Model HVM1 & HSM2 find best Model for Crop Prediction

Figure 1 Model for Proposed Research

The Machine Learning Regression Technique:

Linear Regression is the simplest machine learning techniques used for modeling the relationship between a target (dependent) variable and one or more predictor (independent) variables. Linear regression is easy to understand and implement outlier and multicollinearity show significant impact on model. The model developed by linear regression good for small datasets. The relationship between target variable and predicted variable is shown by equation

$$Y = \beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n+\epsilon$$

Where:

- Y: Target (dependent) variable
- x: predictor (Independent) variable
- β_0: Intercept (the value of y when x is 0)
- β_1: Slope (the change in y for a one-unit change in x)
- ε: Error term (residuals — captures the difference between the actual and predicted values)

Random Forest Regression: It is an extension of the Random Forest algorithm. The random subset of data points, features are used during the tree-building process hence called Random Forest. Random Forest Regression is an ensemble learning method that combines multiple decision trees to improve model stability, accuracy. Random forests reduce the variance and overfitting by combining multiple trees. It Improved Accuracy by averaging predictions from multiple trees. The random forests can capture non-linear, complex relationships in the data. It is more robust to outliers

$$\hat{y}\text{RF} = \widehat{1/\text{T}} + \sum_{t=0}^{1} \hat{y}t$$

Where:

- T - The number of trees in the forest.
- $\hat{y}t$ - The prediction from the t^{th} tree.
- $\hat{y}RF$ - The final prediction of the random forest.

Gradient Boosting Regression: This is the powerful machine learning regression algorithm used for predicting continuous values. It is an ensemble learning technique combines multiple models to create a stronger overall model. The Gradient Boosting Regression builds multiple trees independent and sequentially. The error made previous tree has been corrected by new tree. Gradient Boosting identifies non-linear, complex relationships in the data. It has high predictive power and handles numerical as well categorical data efficiently. As compare to other regression model Gradient boosting is less sensitive to outliers.

Initial Model (Base Model):

$$f_o(x) = 1/N \sum_{i=1}^{N} y_i$$

Where:

- $f_o(x\ )$ is the initial prediction for all instances.
- y_i is the actual target value for the ith training sample.
- N is the total number of samples.

Iterative Updates: At each iteration m, we fit a new decision tree h_m(x) to the residuals, which is the difference between the actual target y_i and the current model's predictions f_m(x)

$$r_i\text{=}y_i\text{-}f_{m-1}(x_i)$$

The model then tries to predict these residuals using the new weak learner (tree):

$$h_m(\text{x})\approx r_i\text{=}y_i - f_m - 1(x_i)$$

Update the Model: The model is updated by adding a fraction (controlled by the learning rate η) of the new tree's prediction to the current prediction

$$F_m(x) = F_m - 1(x) + \eta \cdot h_m(x)$$

Where:

- F_m (x) is the updated prediction after the mth iteration.
- η is the learning rate (a small positive constant, typically between 0 and 1).
- $h_m(x)$ is the output of the mth tree (residual prediction).

Final Model: After M iterations, the final prediction is given by:

$$F_m(x) = F_0(x) + \sum_{M=1}^{M} \eta \cdot h_m(x)$$

Ensemble Learning By combining the strengths of multiple regression models powerful hybrid model can be developed, Ensemble learning technique of machine learning combines the predictions of various base models to build stronger and more reliable model to reduce errors, to produce better predictions and improve performance. Following are the various types of ensemble learning techniques.

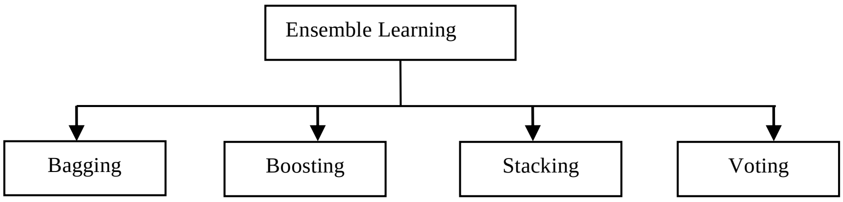

Figure 2 The various Ensemble Learning Techniques

Bagging: Bagging technique trains multiple models by different random subsets with replacement of the sample data. The output of the models is averaged to make final model. It reduces overfitting and variance.

Boosting: Boosting approach combines weak models sequentially, where every new model corrects the mistakes made by the previous model with weight. The final model is a weighted combination of all the models. The model developed by Boosting works well for both classification and regression problems and achieves higher accuracy.

Stacking: In Stacking multiple models are trained by different types of algorithms. The models trained by different algorithms are known as base models. Their predictions are combined and used as input to train new model, called a Meta model. By leveraging the diversity of different models the prediction of Meta model always more accurate. The model developed by stacking can be more flexible than bagging or boosting. Stacking can be implemented using StackingRegressor from scikit-learn library in Python.

Voting: The Voting regression is a powerful ensemble technique used to improve the performance of regression models. It has been combined the predictions of diverse base models which have been trained by same sample dataset. The models used in voting should be independent and must having accuracy more than 50% to built strong model. The training dataset is made by selecting random data. For the classifier the voting has two types hard and soft voting, for regression the final model is made by averaging the predictions of all base models. Voting is easy to implement and require less tuning; often improves accuracy over individual models. The voting regression model can be implemented using the VotingRegressor class in scikit-learn.

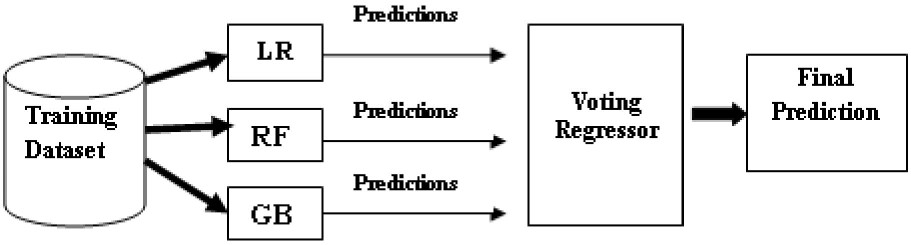

Figure 3 Ensemble Model HVM1

Stacking: The Stacking is an ensemble machine learning technique, also known as Stacked Generalization purposely used for developing powerful hybrid model. The Stacking regression technique combines strengths of multiple predictive models for building hybrid model. The hybrid model improves the overall performance as well as makes more accurate predictions of a regression task. The stacking technique used the predictions of several base models known as level-0 models and trains a meta-model known as level-1 model which produces the final output. Stacking can be implemented using StackingRegressor from scikit-learn library in Python.

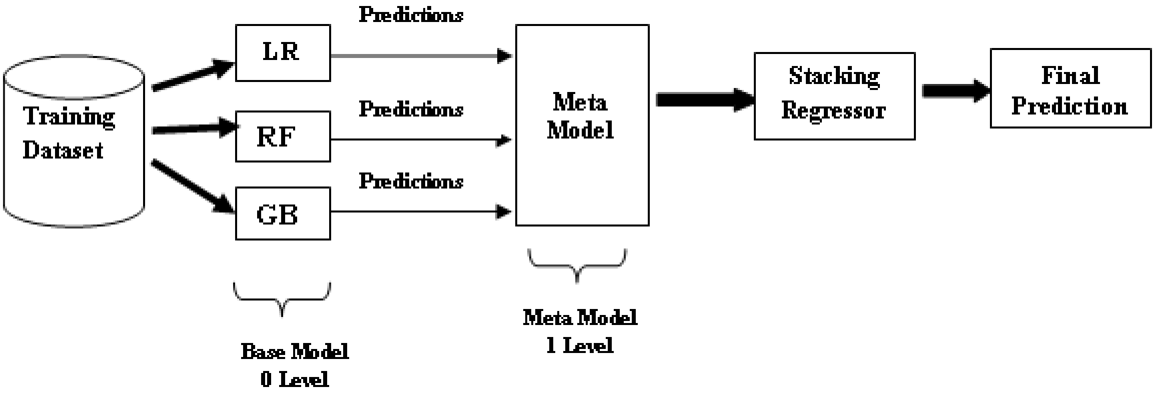

Figure 4 Ensemble Model HSM2

4. Result and Discussion

The trained models developed using various regression and ensemble methods HVM1 and HSM2 have been tested by 20 percent unseen test data for both soil nutrients and environmental factors for particular crop. The performance of the models has been evaluated by calculating R2 Score, MSE, MAE and RMSE.

4.1 The Comparison of Performance metrics Evaluated for various Regression Model

Regression Model	Training Accuracy				Testing Accuracy			
	R2 Score	MAE	MSE	RMSE	R2 Score	MAE	MSE	RMSE
LR	0.946	0.211	2.160	1.470	0.893	0.446	4.536	2.130

RF	0.997	0.103	0.128	0.357	0.943	0.441	2.398	1.549
GBT	0.934	1.084	2.610	1.615	0.905	1.371	4.013	2.003
HVM1	0.888	0.196	0.138	2.123	0.840	0.512	2.331	2.542
HSM2	0.997	1.676	4.509	0.371	0.945	1.911	6.464	1.527

Table 1 The Evaluated Performance of Regression Models

The Table 1 shows the result of performance metrics calculated for regression models consider and developed using ensemble Stacking (HSM2) and Voting (HVM1) techniques. The ensemble HSM2 has lowest MAE, MSE, RMSE and highest R2 Score i.e. 99.7% accuracy for training and 94.5% accuracy for test dataset.

4.2 The Comparison of RMSE evaluated for various Regression and Hybrid model

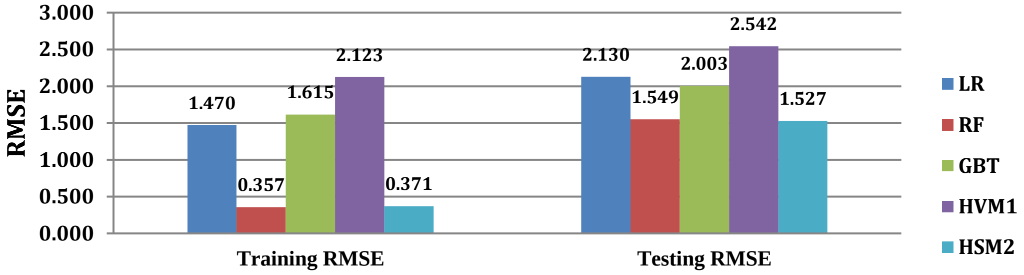

Figure 5 The Comparison of RMSE

The Figure 5 shows the RMSE has been evaluated for all regression models and hybrid models, The HSM2 model performed best among all the models with lowest Root Mean Square Error i.e. 0.371 for training data and 1.527 for test data.

4.3 The Comparison of MSE evaluated for various Regression and Hybrid model

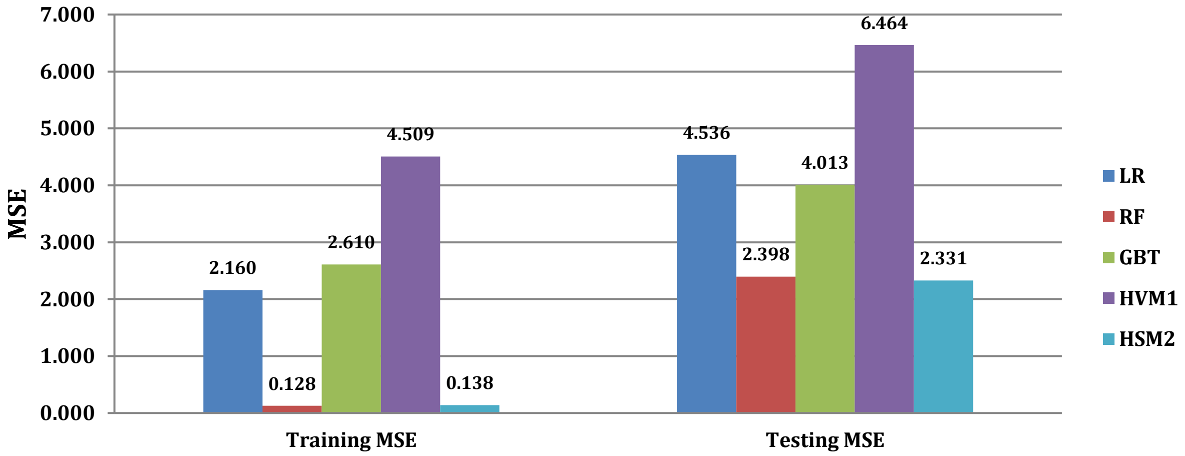

Figure 6 The Comparison of MSE

The Figure 6 shows the Mean Square Error (MSE) has been evaluated for all regression models as well as hybrid models HVM1 and HSM2. The performance metrics MSE calculated for HSM2 model is lowest among all the models i.e. 0.138 for training data and 2.331 for test data.

4.4 The Comparison of MAE evaluated for various Regression and Hybrid model

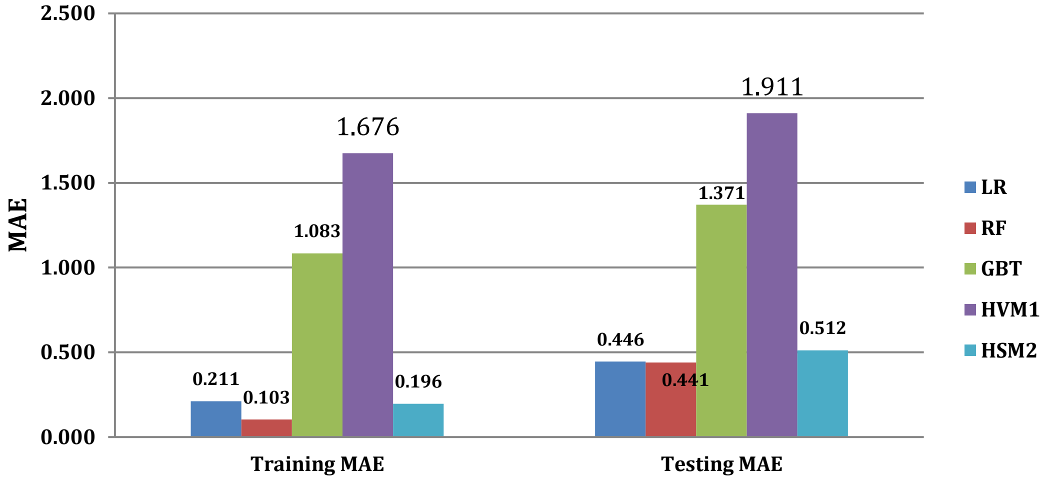

Figure 7 The Comparison of MAE

The Figure 7 shows the Mean Absolute Error (MAE) has been evaluated for all regression models and both HVM1 and HSM2 hybrid models. The performance metrics MAE calculated for HSM2 model is lowest among all the models i.e. 0.196 for training data and 0.512 for test data.

4.5 The Comparison of R2 Score evaluated for various Regression model

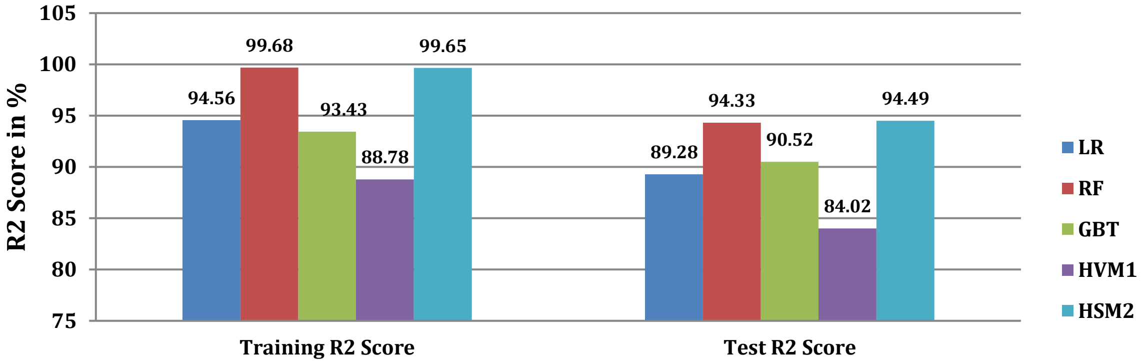

Figure 8 The Comparison of R2 Score

The Figure 8 shows the R2 Score has been evaluated for all regression models and hybrid models, The HSM2 model performed best among all the models with 99.65 % accuracy for training data and 94.49% accuracy for test data.

5. Conclusion

The Agriculture can be revolutionizing by adopting Machine learning (ML) techniques. The ML techniques provide more efficient and sustainable farming practices. The various ML techniques are used to find meaning full insights from historical agricultural data. In this study Linear regression, Random Forest Regression, Gradient Boosting and ensemble (Hybrid) models HVM1 and HSM2 are developed using Voting and Stacking methods respectively to predict most suitable crop for cultivation for soil nutrient, soil pH level as well as environmental parameter of specific area. The ensemble (Hybrid) HSM2 model developed using Stacking technique shown excellent performance in capturing complex patterns in the data and achieved highest R2 Score i.e. 99.65% accuracy for training dataset and 94.49% accuracy for test dataset.

6. Future Scope

In future more crops and parameters will be considered for prediction of best crop for cultivation.

7. References

[1] Ansarifar, Javad, Lizhi Wang, and Sotirios V. Archontoulis. "An interaction regression model for crop yield prediction." *Scientific reports* 11, no. 1 (2021): 1-14.

[2] Bali, Nishu, and Anshu Singla. "Emerging trends in machine learning to predict crop yield and study its influential factors: A survey." *Archives of computational methods in engineering* 29, no. 1 (2022): 95-112.

[3] Ip, Ryan HL, Li-Minn Ang, Kah Phooi Seng, J. C. Broster, and J. E. Pratley. "Big data and machine learning for crop protection." *Computers and Electronics in Agriculture* 151 (2018): 376-383.

[4] Keerthana, Mummaleti, K. J. M. Meghana, Siginamsetty Pravallika, and Modepalli Kavitha. "An ensemble algorithm for crop yield prediction." In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 963-970. IEEE, 2021.

[5] Medar, Ramesh, Vijay S. Rajpurohit, and Shweta Shweta. "Crop yield prediction using machine learning techniques." In *2019 IEEE 5th international conference for convergence in technology (I2CT)*, pp. 1-5. IEEE, 2019.

[6] Nischitha, K., Dhanush Vishwakarma, Mahendra N. Ashwini, and M. R. Manjuraju. "Crop prediction using machine learning approaches." *International Journal of Engineering Research & Technology (IJERT)* 9, no. 08 (2020): 23-26.

[7] Paudel, Dilli, Hendrik Boogaard, Allard de Wit, Sander Janssen, Sjoukje Osinga, Christos Pylianidis, and Ioannis N. Athanasiadis. "Machine learning for large-scale crop yield forecasting." *Agricultural Systems* 187 (2021): 103016.

[8] Pantazi, Xanthoula Eirini, Dimitrios Moshou, Thomas Alexandridis, Rebecca Louise Whetton, and Abdul Mounem Mouazen. "Wheat yield prediction using machine learning and advanced sensing techniques." *Computers and electronics in agriculture* 121 (2016): 57-65.

[9] Rashid, Mamunur, Bifta Sama Bari, Yusri Yusup, Mohamad Anuar Kamaruddin, and Nuzhat Khan. "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction." *IEEE access* 9 (2021): 63406-63439.

[10] Rezk, Nermeen Gamal, Ezz El-Din Hemdan, Abdel-Fattah Attia, Ayman El-Sayed, and Mohamed A. El-Rashidy. "An efficient IoT based smart farming system using machine learning algorithms." *Multimedia Tools and Applications* 80 (2021): 773-797.

[11] Sharma, Bhawana, Lokesh Sharma, Chhagan Lal, and Satyabrata Roy. "Explainable Artificial intelligence for intrusion detection in IoT networks: A deep learning based approach." *Expert Systems with Applications* 238 (2024): 121751.