How to Cite:
Kale, S. S., Patil, P., & Mamatha, G. (2022). Application of stacking ensemble technique in agriculture for prediction of crop yield. *International Journal of Health Sciences*, *6*(S2), 14750–14755. https://doi.org/10.53730/ijhs.v6nS2.8888

Application of stacking ensemble technique in agriculture for prediction of crop yield

Shivani S. Kale
Research Scholar, VTU.Assistant Professor, Dept of CSE, KITCOEK, Kolhapur, Maharashtra

Dr. Preeti Patil
Professor, Department of Information science and Engineering, RNS Institute of Technology, Bengaluru. 560098

Dr. Mamatha G.
Professor & HOD, Department of Information Technology, DYPCOE, Akurdi, Pune. MH

***Abstract*---**In comparison to other methodologies, Machine Learning has proven to be an effective tool for making reasonable predictions. Initially, a single machine learning model is used on data to make predictions; however, instead of using a single machine learning model, a collection of ML models is employed instead. Models can reduce prediction variance, bias, or both. The result provided will be more appealing and correct if appropriate features affecting the prediction values are supplied. This study presents a framework for selecting feature subsets using the L2 regularization feature selection technique. The study took into account 11 crops and 15 characteristics from 300 districts across India. For predicting yield, basic learners' models such as Linear Regression, LASSO Regression, Decision Trees, Random Forests, Extreme Gradient Boosting (XGBoost), and Support Vector Machines are initially trained and tested. Base learners are used to create the stacking ensemble and the proposed optimum stacking ensemble. In terms of the performance measures employed, the proposed Optimized stacking ensemble shows approximately (MSE =6408, Accuracy=85%) outperforms the Normal stacking ensemble (MSE=8315, Accuracy=83%).

***Keywords*---**stacking ensemble, prediction, accuracy.

Manuscript submitted: 09 March 2022, Manuscript revised: 18 April 2022, Accepted for publication: 27 May 2022

Introduction

The Paper outlines study and implementation of different ML algorithms on variety of crops using variables which includes temperature, rainfall and different soil parameters for tentatively 300 districts of INDIA. The objective of paper is to suggest farmer what will be the tentative yield of crop. So, in first step base learner model is designed and applied and performance is measured. In step 2 ensemble technique is applied and measured the performance. Step 3 Optimized ensemble is applied and performance metrics are calculated. Comparison of all the models depending on different parameters are carried out. The Paper is basically organized into four sections. Dataset and Data preprocessing is included in section 1. The ML methodologies and proposed stacking Ensemble technique are described in Section II. Section III is devoted for discussion of performance of models with the help of results obtained. To finish, the paper concludes with Section IV.

Data set

The dataset includes 15 features namely state, District, year, season, crop, temperature, rainfall, production, N, P, K, Ph, Fc, Oc, Zn. The dataset is having nearly 2 lakh records. The data is from year 1997 to 2014.The data is collected from government website.

Table 1: Dataset resources

Dataset resource Name	Link	Parameters
Open Government Data Platform INDIA	https://data.gov.in/	state, District, year, season, crop, state, District, year, season, crop
Department of Agriculture, Cooperation and Farmer's Welfare	https://agricoop.nic.in/en	N, P, K, Ph, Fc, Oc, Zn
World Bank Group	https://climateknowledgeportal.worldbank.org/download-data	temperature, rainfall

Features Selection

Figure 1: Feature Selection

Machine learning Methodologies

Model 1: Linear Regression Model 2: Lasso Regression Model 3: Decision Trees
Model 4: Extreme Gradient Boosting (XGBoost)
Model 5: Support Vector Machines with polynomial kernel Model 6: Random Forest Model 7: Stacking Ensemble Technique
Model 8-: Proposed Stacking ensemble using L2 Regularization for feature selection

Proposed research work

The Proposed work uses L2 Regularization for feature selection, the predictions obtained by feature selection methods are compared to choose the best feature selection techniques and best prediction accuracy. Stacking ensemble with L2 regularization applied. The proposed methodology used in this research paper is described in Fig. 5

$$\mathrm{J(W;X,y)} = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

Equation 1: L2 regularization

M= number of samples, P= Number of features, x_{ij}= i^{th} sample j^{th} feature, y_i=target value for i^{th} sample, wj=weight of j^{th} feature input from previous layer, λ=It is regularization parameter used to control the selected values of w by optimization algorithm. The regularization parameter or penalty puts constraint on the coefficients *(w)*. The penalty term (lambda) regularizes the coefficients such that if the coefficients take large values the optimization function is penalized. So, L2 regularization or ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity.

Proposed Optimized stacking ensemble algorithm

Input: Training data (X, Y). Trained base models: M1=M 6*20, M2=M 6*10
Output: Trained meta learner, P

```
For Layer 1
Initialize x= [ ]
for m in M1 do
p = m. predict(X)
x. append(p)
end for
Z-hat = x. L2 Regularization
For Layer 2
Input: [X, Z-hat]
Initialize x1= []
For m in M2 do
p = m. predict(Z-hat)
x1.append(p)
end for
z1-hat:x1.L2 regularization
Params= z1-hat.L2 Regularization parameters
results = gridsearch. bestfit (x1, z1-hat, params)
P= results. best_model
return P
```

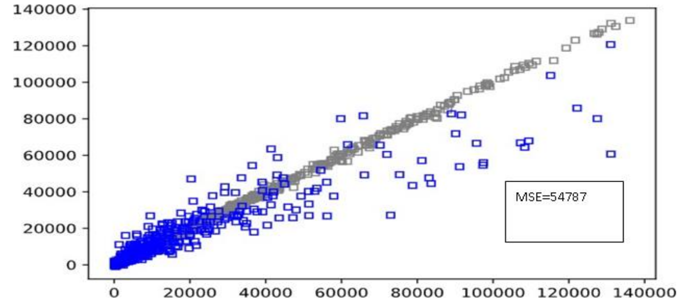
Algorithm 1: stacking ensemble technique using L2 regularization

Results

Models	Mean absolute error	Mean squared error	Median absolute error	Explain variance score	R2 score
Stacked ensemble Regression	36125.70633	54787.3162	21440.69531	0.93578	0.93578
Linear Regression	58370.57949	78868.06705	41386.2901	0.86692	0.86692
Lasso Regression	58370.30866	78867.97076	41386.0682	0.86692	0.86692
Decision Tree Regression	50847.20812	74986.08297	30755.84722	0.87977	0.8797
Random Forest Regression	35561.24348	54152.06073	21779.544	0.89468	0.88533
XGBoost Regression	39693.18455	56601.46108	25132.275	0.88709	0.87785
Support Vector Machines Regression	164878.81372	224202.20273	126421.57047	0.00447	-0.07546

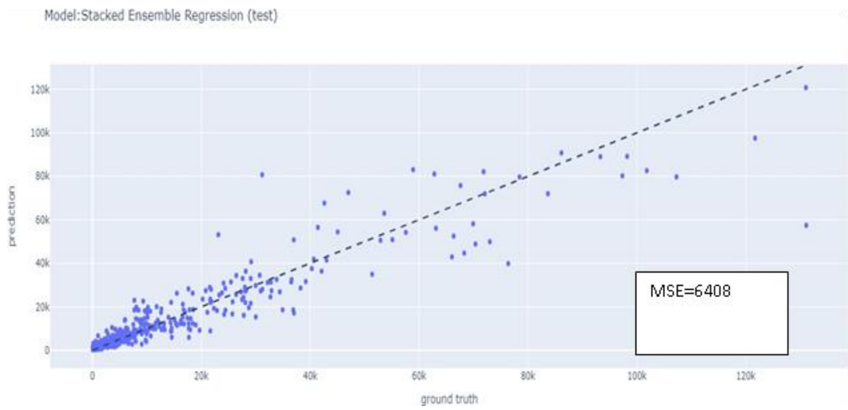
Table 1: Comparison of performance evaluation of stacking ensemble with base model for wheat crop

Prediction by Stacking Ensemble



Graph 1: Predicted Versus Actual Values for stacked Ensemble

Proposed Optimized stacking ensemble



Graph 2: Predicted Versus Actual Values for Optimized stacked Ensemble

Conclusion

Tentative prediction of yield will help the farmer regarding decision about cultivation of crops. Different base models are used to create ensemble model. The optimized ensemble model using L2 regularization gives better performance compared to stacked ensemble. Comparison of results of different base models with proposed model is carried out in the study. The proposed optimized stacking ensemble model that tries to balance both bias and variance of predictions and applies functions to discover the optimal features to get improved predictions. The Proposed model gives 3% higher accuracy than normal stacking Ensemble technique.

References

1. Mrs. Shivani S. Kale and Dr. Preeti Patil, “Data mining technology with fuzzy logic, neural networks and machine learning for agriculture”, at ASIC book series “Data management, analytics and innovation by springer 2018.
2. Mr. V. Sellam and E. Poovammal,” Prediction of crop yield using regression analysis”, Indian Journal of Science and technology,Vol 9 (38),October 2016.

3. Mrs. Shivani S. Kale and Dr. Preeti S. Patil," Use of data mining technology in agriculture sustainable development", IJCRT, Volume 6, Issue 2. April 2018.
4. 4.Sanjay D. Sawaitul, Prof. K.P. Wagh, Dr. P.N. Chatur, 2012. Classification and Prediction of Future Weather by using Back Propagation Algorithm An Approach. International Journal of Emerging Technology and Advanced Engineering 2:110-113.
5. 5.Tripathi, S., Srinivas, V. V., & Nanjundiah, R. S. 2006.Downscaling of precipitation for climate change scenarios: a support vector machine approach. Journal of Hydrology -330:621-640.
6. Altannar Chinchulunn, Petros Xanthopoulos, Vera Tomaino, P.M.Pardalos., January-June 2010. Data Mining Techniques in Agricultural and Environmental Sciences. International Journal of Agricultural and Environmental Information Systems 1:26-40.
7. Abraham. 2005. Artificial Neural Networks: Handbook of Measuring System Design, John Wiley & Sons.0:470-02143-8.
8. B. Ji.2007. Artificial neural networks for rice yield prediction in mountainous regions.Journal of Agricultural Science 145:249–261.
9. B.A. Smith. G. Hoogenboom and R.W. McClendon. 2009.Artificial Neural Networks for Automated Year round Temperature Prediction.Computers and Electronics in Agriculture 68:52–6.
10. M. Schaap and W. Bouten. 1996. Modeling water retention curves of sandy soils using neural networks. Water Resour. Res.32: 3033–3040.
11. S.K. Starrett and G.L. Adams. 1997.Using artificial neural networks and regression to predict percentage of applied nitrogen leached under turfgrass. Commun. Soil Sci. Plant Anal 28:497–507.
12. Rinartha, K., & Suryasa, W. (2017). Comparative study for better result on query suggestion of article searching with MySQL pattern matching and Jaccard similarity. In *2017 5th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-4). IEEE.
13. Rinartha, K., Suryasa, W., & Kartika, L. G. S. (2018). Comparative Analysis of String Similarity on Dynamic Query Suggestions. In *2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)* (pp. 399-404). IEEE.