Journal of Theoretical and Applied Information Technology
31st December 2023. Vol.101. No 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ENSEMBLE APPROACH FOR HARVEST VINTAGE FORECAST WITH MACHINE LEARNING TECHNIQUES: AN EXPERIMENTAL STUDY

[1]DR.SURESH KUMAR PITTALA, [2] DR. KONGARA SRINIVASA RAO [3]BALAJI.TATA, . [4]P.RAVI KUMAR [5]N.JAYA, [6]DR Y ANURADHA [7]KURRA UPENDRA CHOWDARY

[1]Associate Professor, Dept of ECE, R.V.R& J.C.College of Engineering, Guntur, AP,
[2]Dept. of CSE, Faculty of Science& Technology, ICFAI Foundation for Higher Education Hyderabad,
[3]Sr.Asst Professor, Dept of ECE, PVP Siddhartha Institute of Technology, Vijayawada,
[4]Assoc Professor, Dept of ECE, Shri Vishnu Engineering College for Women (A),Bhimavaram,
[5]Professor, Department of EIE, Faculty of Engineering and Technology, Annamalai University,
[6]Associate Professor, CSE Dept, GVP College of Engineering, Visakhapatnam
[7]Associate Professor, Dept of ECE, R.V.R& J.C.College of Engineering Guntur,
E-mail: dr.sureshkumarpittala@gmail.com drksrao@ifheindia.org balajitata@pvpsiddhartha.ac.in
ravikumar_tnk@svecw.edu.in jayanavaneethan@rediffmail.com
anuradhayarlagadda@gvpce.ac.in kupendra@rvrjc.ac.in

ABSTRACT

Since crop yield forecasting directly affects the production and security of food, it is an essential part of agricultural research and development. The traditional methods of calculating harvest vintages centered on agriculturalists' explanations and expertise have become increasingly challenging as a result of the rapid ups and downs in soil and climatic state of affairs. Machine-learning techniques have been applied in recent years to find a solution to this issue. This study centers around the utilization of a few AI models, for example, nearest neighbor relapse, closest polynomial relapse, irregular woodland relapse, slope helped tree relapse, and backing vector relapse, for the expectation of farming efficiency. The raw data was transformed into a format that is conducive to machine learning using efficient feature selection techniques. The study's findings showed that, in comparison to other tactics, when choosing features, using an ensemble technique can lead to more accurate predictions. By combining several data sources and machine learning algorithms, the ensemble approach generates a detailed and precise forecast. This work highlights both the benefits of using machine learning techniques to forecast agricultural amount produced and the recompenses of using a collaborative line of attack to feature selection

Keywords: *Harvest, Vintage, Ensemble, Forecast, Agriculture*.

1. INTRODUCTION:

Crop forecasting is a vital area of research for the agricultural sector. This is due to the enormous effects that a number of environmental conditions, counting soil, temperature, dampness, and precipitation, have on harvest vintages [1]. Farmers used to be able to make decisions about which crops to grow, how to monitor their progress, and when they may be harvested based on their observations and expertise. However, because of the recent rapid changes in the environment, the unindustrialized communal has found it supplementary difficult to be dependent on on these unadventurous approaches [2].

Therefore, machine learning technologies have been pragmatic to aid in harvest likelihood. Crop yield prediction, which ascertains the likely output of a certain crop in a particular place, is a crucial aspect of agriculture. Accurate crop production estimates may help farmers, agribusinesses, and governmental organizations make informed decisions about crop planting, harvesting, and marketing [3]. The importance of agricultural production forecasts is illustrated by the following instances: The structure continually recommends examination areas based on the uncertainties of the statements, which are prepared, before the error is revealed. according to the details of how the experiment was implemented and the developer's feedback from prior checkpoints. Better resource management: By accurately forecasting agricultural yield, farmers can manage their resources more effectively. This may entail deciding on the best crop to produce, the perfect

proportion of compost, and the best planting and gathering windows.

Healthier planning and groundwork: By employing realistic crop production estimates to ensure they have the necessary infrastructure and resources, ranchers can design and be ready for the collect ahead of time. Getting ready for the yield's showcasing, stockpiling, and transportation might be fundamental.

Higher-quality decision-making Agribusinesses and governmental entities may make better decisions on the acquisition, sale, and importation of harvests with the help of accurate crop production forecasts. They might also use the data to plan ahead and ensure that there is a consistent supply of food for the populace.
Increased food security: Since an adequate quantity of food can be secured, food security may be increased by accurately anticipating agricultural production. It very well may have the option to keep away from food deficiencies, value climbs, and food squander by doing this. Precise horticultural result evaluations might lessen the misuse of assets like water, manure, and land. This leads to better resource management. With this information, farmers may make better use of these resources, leading to more sustainable farming practices [5].

2. LITERATURE SURVEY:

Crop production forecasting is a challenging task that calls for several measurable and numerical techniques that are constantly being enhanced [6]. This approach could be advantageous for product creation, design, and improvement. To lead factual examination, which is used to make decisions about assorted occasions and direct financial navigation, numerical data is required. According to Murithi, the greater the level of understanding of the phenomenon, the more accurate the information and judgments that can be produced, and the more accurate the numerical data. One of the essential issues in the calm temperature zone is deciding the impacts of agro-climatic circumstances on the creation of winter crops, quite grains.
The number and recurrence of days with temperatures above 5°Cduring the colder time of year, as well as the quantity of days with temperatures above 0°C and 5°C, are pivotal elements that decide the yield of winter crops
Agro meteorological characteristics change, thus being able to predict them properly is also essential for precise production forecast.

The study also found that predicted profitability for lupine production under climate change is greater than that seen between 1990 and 2008. The HadCM3 scenario was shown to be the most favorable of the three models [8]. They utilised images from RadarSat-2 and Sentinel-1 to differentiate between distinct crops.

3. METHODS:

Methodology

Transparency, repeatability, and correctness of the study may all be considerably enhanced by a well-written methods section. A key component of the research is the approach used in agricultural yield prediction systems [11]. If other researchers can understand the prediction system better, they will be able to build upon the earlier study. The following is the study's suggested methodology:
Data collection: Data collection is the process of obtaining information in order to achieve a certain objective. To create a ML Model for harvest prediction, it is necessary to get soil, meteorological, and crop information.

Preprocessing entails placing raw data in a manner suitable for analysis. By replacing any missing or null values, the data is cleaned in this procedure and put into a comprehensible manner. The preparation dataset is utilized to prepare AI calculations and give precise forecasts, while the testing dataset is utilized to assess the presentation of the calculation. [10]
Numerous factors affect the production and yield of crops. These elements, sometimes known as "characteristics," aid in predicting the production of a particular crop during the course of the year.

To determine which machine learning algorithm is the most effective in predicting crop yield, several must be compared and examined. For this specific dataset, this study investigates the efficacy of several distinct methods [11].

Transparency, repeatability, and correctness of the study may all be considerably enhanced by a well-written methods section. A key component of the research is the approach used in agricultural yield prediction systems [12].

Training and Evaluation:

Model preparation is the method involved with educating an AI model to make forecasts in light of information. The target of model preparation in crop forecast is to make an AI framework that can precisely gauge crop creation utilizing soil, climate, and yield information.

Model assessment is the process of assessing a machine learning model's effectiveness. Finding out how well a crop prediction model can predict crop yield based on input data is the aim of model evaluation[14].
Model refining is the process of enhancing or altering a machine learning model to improve performance.

Deployment:

When a machine learning model is used and placed into production, it is said to have been deployed.

The nonstop assessment of a conveyed AI model's presentation is alluded to as "model checking."

The flowchart addresses the progression of the yield estimating arrangement of various stages.

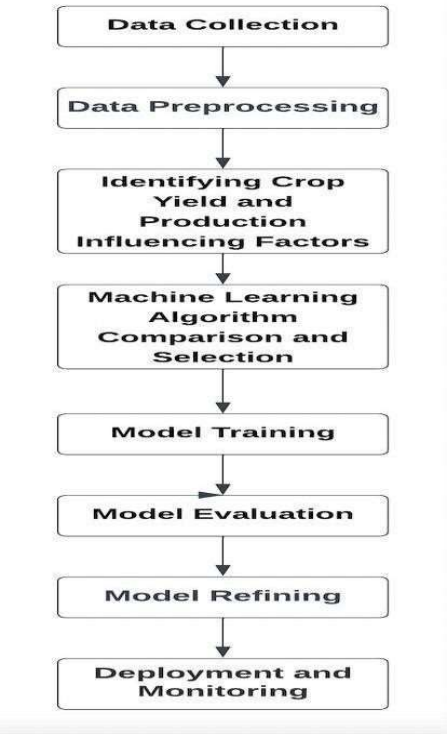

Fig 2. Algorithm Development

4. ALGORITHMS:

- **Linear Regression**

Using supervised machine learning methods like linear regression, it is possible to model the relationship between a dependent variable and one or more independent variables. Finding the ideal linear connection, which is a linear combination of the independent variables as their coefficients, is the goal.

$$y= a0+a1x+ \varepsilon \text{ -----(1)}$$

- **Polynomial Regression**

Polynomial relapse is utilized to demonstrate the connection between a reliant variable and at least one free factors. At the point when there is a non-direct connection between these factors, it is utilized. This kind of relapse involves a polynomial condition as the model, and the maximal force of the free factor is bigger than one.

Random Forest Regression

To address regression concerns, machine learning algorithms like random forest regression integrate the predictions of many decision trees. As opposed to a solitary choice tree, which just uses the result from one tree to make expectations, Irregular Woodland Relapse integrates the results from a few trees.

Gradient BTR
Gradient-BTR is one of the machine learning methods for tackling regression problems. It works by combining the predictions from different decision trees. This technique builds a sequence of trees, each of which aims to correct the errors made by the one before it. Conventional relapse calculations construct a tree each in turn.

Prediction equation: y(pred) = y1 + (eta * r1) + (eta * r2) +.. + (eta * rn)

Nearest Neighbor Regression

The "nearest neighbour regression" technique, which is employed in machine learning, finds the data points that are closest to a new observation and forecasts it using their average values. You can alter the number of closest neighbours.

SVM

Relapse related issues are managed utilizing the AI approach Backing Vector Relapse (SVR). Support vector machines, which are regularly utilized for arrangement, structure the underpinning of this framework. The objective of SVR is to find the line

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

or hyper plane that offers the best edge and partition between the objective qualities and the gauges.

5. TRAINING:

The following stages can be done when the information preprocessing and model determination are finished

Data Split Up

In the code you gave, the train-test split capability from the scikit-learn bundle is utilized to tell the best way to direct a train-test split. The objective of the train-test split is to separate a dataset into a preparation set and a test set. The train testsplit capability takes as sources of info highlights X, target Y, the test size (testsize=0.3), and the test size. As indicated by this, 30% of the information will be utilized for testing while the leftover 70% would for train. The arbitrary state choice is utilized to lay out the irregular seed (irregular state=42), guaranteeing that a similar train-test split will be utilized each time the code is run..

Holdout method

The holdout strategy is a well-known machine learning technique for data separation [15]. This method uses a portion of the available data as the validation set and the rest data as the training set to assess the performance of the model. The ratios between the training and validation sets might range from 70/30 to 80/20 and are often determined at random. Although the holdout approach is straightforward and simple to implement, the unexpected data splitting might result in significant variations in the model's performance. To move beyond this limitation, cross-approval draws near, for example, k-overlap cross-approval, have been created to improve the information parting process and give more exact model appraisal.

Cross-validation

The cross-fold method of data splitting may be used to break up a large dataset into smaller, more manageable chunks. Using this method, the dataset is divided into training and validation sets.The massive dataset is divided into k equal-sized subgroups using a positive integer, k. After every subset has been utilized once as an approval set, the leftover k-1 subsets are then utilized for preparing [15]. Throughout the span of this strategy, which is done k times, every information point is utilized for preparing k-multiple times and approval once, with every subset going about as the approval set once.

Training models

In this exploration, we prepared different models. The trained models and their means are as per the following:

- Linear Regression.
- Polynomial Regression:
- Random Forest Regression:
- Gradient Boosted Regression
- Nearest Neighbor Regression:.
- Support Vector Regression:

Evaluation Metrics

AI utilizes assessment measurements to evaluate the viability of a model. In spite of the fact that there are a few evaluation estimates open, probably the most frequently utilized measurements incorporate exactness, standard deviation, cross-approval score, and preparing score [15]. Exactness is an ordinarily involved assessment metric in AI. All it estimates the extent of a model's exact forecasts to its expectations. The scope of a gathering of insights is shown by the standard deviation. In AI, the standard deviation is an estimation of the exhibition changeability of a model. It gives data on the deviation a model's exactness makes from the mean..

A statistic for measuring how well a model generalises to new data is the cross-validation score. Partitioning the training data into several subgroups and training and validating the model on different subsets is known as cross-validation. This procedure is repeated multiple times, and the cross-validation score is determined as the model's average accuracy over all repetitions [15]. The training score of a model serves as a measure of how well it fits the training data. Using the training set of data, the model's accuracy is assessed. A high training score is frequently indicative of an overfitting model, whereas a low training score is typically indicative of an under fitting model.. All in all, these measurements are much of the time utilized in AI to assess the adequacy of models. The mean, standard deviation, cross-approval score, and preparing score are a few measurements that might be utilized to survey a model's exactness and generalizability.

6. RESULTS AND DISCUSSION

Six unmistakable AI models, including closest neighbor, support vector relapse, irregular woodland, and straight relapse with L2 regularization, were utilized in the review to gauge rural creation. The consequences of the examination showed that the exactness of each

model fluctuated. Straight relapse and direct relapse with L2 regularization had comparable exactness of 0.44, demonstrating that L2 regularization did not improve the performance of linear regression for this particular dataset. Polynomial regression fared better, with an accuracy of 0.68, recommending that a more intricate model may be more reasonable for this utilization. Be that as it may, the model's intricacy ought to be considered since overfitting could happen assume the model is excessively complicated. Arbitrary timberland and closest neighbor seem, by all accounts, to be the most dependable models for anticipating agrarian creation, with exactness appraisals of 0.83 and 0.81, separately. Support vector relapse's exactness was 0.59, which was less precise than the precision of different models.

Table 6.1 Learning Curve For The GBT.

Training Examples	CV score	Training score
50	0.5	1
250	0.7	0.9
600	0.75	0.85
850	0.79	0.85
1100	0.8	0.85

To determine which model predicts crop yields the best and how well it performs in relation to the others, the learning curves of several models may also be compared. The learningcurves, which also provide visual representations of how well various machine learning models predict crop yields, may confirm the study's conclusions.The results of this study show the benefits of feature selection when using an ensemble approach to estimate crop yield. The ensemble approach brings together many models to get a more precise forecast. This study's ensemble method yielded a prediction accuracy of 0.89, which was higher than the individual prediction accuracy of each machine learning model used.

Table 6.2 Learning Curve For The Random Forest.

Training Examples	CV score	Training score
50	0.5	1
250	0.6	0.8
600	0.85	0.95
850	0.89	0.99
1100	0.88	0.915

Table6.3 Learning Curve For The Nearest Neighbor

Training Examples	CV score	Training score
50	0.5	1
250	0.6	0.9
600	0.875	0.82
850	0.89	0.86
1100	0.812	0.87

Table 6.4. Learning Curve For The SVM

Training Examples	CV score	Training score
50	0.5	1
250	0.7	0.9
600	0.75	0.85
850	0.79	0.85
1100	0.8	0.85

7. INFERENCE AND IMMINENT SCOPE:

The outcomes demonstrate the way that include determination with a gathering strategy can create forecasts that are more exact than those acquired with different methodologies. An extensive and exact estimate is delivered utilizing an outfit strategy, which joins many models. This approach gauges the advantages and downsides of each unmistakable model to give a more precise estimate. This study utilized an outfit approach for highlight determination, which has been shown to produce a more precise gauge of harvest yields than utilizing individual models. This features the significance of component determination in AI for foreseeing rural efficiency as well as the utility of utilizing a gathering technique. The results of this examination show the meaning of element determination and AI model choice for precise yield creation expectation. The outcomes showed that the closest neighbor and arbitrary backwoods models were the most dependable, while the straight and backing vector relapse models performed similarly inadequately. The outfit approach, which consolidated the outcomes from many models, created the most dependable forecast of harvest yield. This features the need of using a scope of models and strategies to work on the precision of yield projections.

All in all, this work gives significant data on how actually unique AI models anticipate crop yields and the advantages of involving an outfit method for highlight determination. As per the discoveries, future examination in this field ought to zero in on further developing yield conjecture precision by executing state of the art AI models and element determination procedures. The examination likewise lays out the establishment for the improvement of more mind boggling and definite rural result expectation frameworks, which will immensely affect food creation and security.

REFERENCES:

[1]. Raza, S. A., A. Raza, and M. A. Hussain. "Crop yield prediction using machine learning techniques: A review." Computers and Electronics in Agriculture 158 (2019): 205-217

[2]. Singh, N., et al. "A comparative study of machine learning algorithms for crop yield prediction." IEEE Access 6 (2018): 11085-11093.

[3]. .Praveen, S. P., Murali Krishna, T. B., Anuradha, C. H., Mandalapu, S. R., Sarala, P., & Sindhura, S. (2022). A robust framework for handling health care information based on machine learning and big data engineering techniques. International Journal of Healthcare Management, 1-18.

[4]. Devi, R., and S. R. Meka. "Crop yield prediction using machine learning techniques: A systematic review." Journal of King Saud University- Computer and Information Sciences 32.4 (2020): 480-490.

[5]. De Bie, C. A. J. M., et al. "Crop yield prediction using machine learning: Considerations for domain adaptation." Remote Sensing 12.11 (2020): 1825.

[6]. Yang, H., et al. "Crop yield prediction using deep learning: A review." Remote Sensing 11.20 (2019): 2405.

[7]. Sai, N. R., Chandana, B. S., Praveen, S. P., & Kumar, S. S. (2021, November). Improving Performance of IDS by using Feature Selection with IG-R. In 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud)(I-SMAC) (pp. 1-8). IEEE.

[8]. Zhang, L., et al. "Crop yield prediction using machine learning models: A systematic review." Plant Phenomics 2021 (2021).

[9]. Zhong, Y., et al. "A comprehensive review of machine learning for crop yield prediction." Frontiers in Plant Science 11 (2020): 525.

[10]. Xia, Y., et al. "A review of machine learning methods for crop yield prediction." Agricultural and Forest Meteorology 288 (2020): 107981.

[11]. Du, X., et al. "Machine learning-based yield prediction of tomato crops in greenhouse production." Computers and Electronics in Agriculture 170 (2020): 105251.

[12]. Xiong, W., et al. "Deep learning for crop yield prediction in precision agriculture: A review." Plant Methods 17.1 (2021): 3.

[13]. Wu, Y., et al. "Crop yield prediction with machine learning and remote sensing data: A review." Information Fusion 67 (2021): 177-189.

[14]. Marrapu, B. V., Raju, K. Y. N., Chowdary, M. J., Vempati, H., & Praveen, S. P. (2022, January). Automating the creation of machine learning algorithms using basic math. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 866-871). IEEE.

[15]. Liao, H., et al. "Crop yield prediction using machine learning and big data: A review." Big Data Mining and Analytics 2.2 (2019):86-94

[16]. Feng, J., et al. "A review of crop yield prediction methods based on machine learning.." Journal of Agricultural Science and Technology (2020):1-15

.