Spectrum of Engineering Sciences

ISSN (e) 3007-3138 (p) 3007-312X

Volume 4, Issue 1, 2026

INTELLIGENT FORECASTING OF AGRICULTURAL PRODUCTION THROUGH ADVANCED DATA MINING ALGORITHMS

Malik Nauman[1], Zain Shaukat[2], Aqib Mehmood[*3], Waqas Ahmed[4], Mubashir Zainoor[5], Salman Ali Khan[6]

[3]aqibmehmood@inu.edu.pk

DOI: https://doi.org/10.5281/zenodo.18219703

Keywords

Crop Yield Prediction, Machine Learning, Data Mining, Decision Tree, Agricultural Analytics, Environmental Factors, Regression Models, Precision Agriculture

Article History

Received: 12 November 2025
Accepted: 24 December 2025
Published: 12 January 2026

Copyright @Author

Corresponding Author: *

Aqib Mehmood

Abstract

Precise estimation of farm production is a crucial component of agricultural planning, food safety, and financial stability, particularly in an agrarian economy like Pakistan's. This paper examines the practice of data mining and machine learning to predict crop yields in different environmental and agronomic conditions. Several predictive models, comparable linear regression, multinomial naive Bayes, decision tree, XGBoost Regressor, Stochastic Gradient Descent (SGD) Regressor, kernel ridge regression, Elastic Net, Bayesian ridge regression, Gradient Boosting Regressor, and Support Vector Regression (SVR), were developed and tested on historical agricultural data that included type of crop, cultivated land, average temperature, rainfall, and pesticide use. Accuracy, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) were also measures of model performance. According to the experimental results, the Decision Tree model is the best in the comparison with the other methods, with an accuracy of 93.27, an MAE of 6.73, and an RMSE of 2.59. The outcomes suggest that decision tree-based techniques are very successful in the prediction of crop yields and could be of use in giving decision support to the farmers and policymakers. A decision tree is an algorithm of managed machine learning, which applies a hierarchical, tree-based framework of decision-making and prediction.

INTRODUCTION

Crop yield is the total production of agriculture that is formed within a certain piece of land within a particular year. It is a powerful indicator, as it assists in assessing food security and comprehending the circumstances that bring about changes in agricultural products in the long run. The efficiency of food production is usually measured by crop yield, as it is the measure of the volume of harvest per acre of land. The Pakistani economy greatly relies on crop production, and therefore, it is often referred to as the backbone of the economy. The agricultural sector supports a significant proportion of the populace and contributes substantially to the country's Gross Domestic Product (GDP). In Pakistan, the principal crops include wheat, cotton, rice, sugarcane, and various food grains. The two major natural resources that sustain agricultural activities in the country are water and cultivable land. Agriculture is a major source of the Pakistani GDP (18.9 percent) and a source of employment for 42.3 percent of its labor force. The most common crops are wheat, cotton, and rice, with a combined summation of over 75 percent of the total value of agricultural produce, although many farmers in Pakistan cannot attain optimum crop production because of a number of factors. Weather conditions, particularly the rainfall patterns, are very reliant on crop production, as this highly determines the growth and development of crops. Farmers require guidance and proper assessment at the right time in such uncertain conditions in order to assist them in enhancing crop productivity. Though farming has been an ancient profession in Pakistan, the output is usually not satisfactory due to the continual interaction of many factors that dictate the yield of crops. In order to produce food for the people with a population of about 210 million, there is a need to enhance the productivity of agriculture. The type of soil, rainfall, temperature, and availability of water are some of the factors that are critical in determining the amount of crops produced. The recent development in information technology and data analysis has presented new research possibilities in the field of agriculture, such as predicting the harvest. Ineffective yield prediction models are one of the biggest challenges that should be addressed with the help of valid and open data sources. The computational methods of data mining offer superior alternatives to the analysis of huge agricultural data and forecasting of the future growth of crops. To investigate the state of agriculture and determine the production of crops in the future, different methods of collecting data are employed. Agricultural data analysis is a developing area of research that aims to derive useful patterns from large datasets that can be used in decision-making.

i. No dependable system exists in place to predict the annual crop yields with accurate predictions, and this restricts our knowledge about the trends of yields.

ii. The aspects of the involvement of numerous environmental and agricultural factors are the major obstacles to the complexity of yield prediction.

iii. To improve the efficiency of crop yield, storage, and transportation, it is essential for farmers and government authorities to receive timely recommendations that facilitate informed decision-making and optimize outcomes.

2. Problem Statement

Crop-yield prediction is quite a problem, as there is still no dependable or uniform method of estimating a yearly yield of any crop based on our knowledge regarding it. The issue of crop yield prediction is still a big challenge, especially when you are dealing with small and sparse sources of data. This type of uncertainty presents a problem to both the farmers and the government establishments, who would want to receive timely and accurate approvals that would assist them in making well-informed decisions regarding how to produce, store, and transport, and the cost of production. This question has crucial importance to the improvement of agricultural planning and food security. With the rising impacts of climate change, unpredictable weather conditions, and the altered relationship between pests and diseases, the developed prediction techniques have never been as significant as they are currently. Indicators: Data-based techniques such as data mining and machine learning can provide actionable information, which the interested parties take positive measures to ensure that they become more productive, reduce their losses, and make maximum use of their resources. The main problems that were identified in the sequence of our research are the following.

i. There are no good ways of predicting the annual crop yield that would provide us with significantly better information on the crop yield trend as it is today.

ii. The problem of complicating yield forecasts is one of the significant issues that should be solved in relation to the desired information.

iii. Farmers and the government need appropriate advice and references on how they can attain higher yields of the crops and store them, transfer costs, etc.

3. Literature Review

The agricultural sector in the modern age of technological advancements and innovation is anticipated to offer an avenue to the farmers of the world to have effective tools and solutions in order to be productive and yield high crop production. The economy of Pakistan is highly based on agricultural growth and production, and a significant number of the population are dependent on farming as their main livelihood. Production of crops is directly associated with the national activity aimed at ensuring food security, or the possibility for farmers to control their production and market it successfully. Having a proper estimation of crop yields is important in helping farmers, policymakers, and other interested parties plan agricultural activities, including harvesting, storing, pricing, and marketing. Different intelligent methods have been implemented in crop yield prediction systems (DSS), such as artificial neural networks (ANN), Bayesian networks, Support Vector Machines (SVM), and other machine learning algorithms. Scholars have investigated various methods of enhancing the precision and trustworthiness of agricultural yield production systems. K. The yield estimation system suggested by Lata and B. Chaudhri [4] is an estimation model that utilizes the data mining algorithm, and it emphasizes the necessity of the analytical approach in agricultural decision-making.

3.1 Crop Yield Prediction

Data mining methods are gradually being used to forecast crop yield and also to determine the appropriate crops to plant in agricultural lands in an attempt to enhance the efficiency and productivity of agricultural lands. These predictive methods utilize agronomic, environmental, and historical agricultural data to improve strategic planning and optimize resource utilization. Nevertheless, a significant portion of the current systems has issues with changing to the dynamic and complex environment of farming and delivering valuable and timely advice to the farmers. Akunuri Manjula and G. Narsimha (2015) suggested a framework named XCYPF: A Flexible and Extensible Framework on Crop Yield Prediction, to overcome these weaknesses. The proposed framework can exploit data mining with the view to attract the accuracy, adaptability, and scalability of yield forecast models in various crops and varying conditions. Research has revealed that regression models, decision trees, neural networks, support vector machines, and additional statistical data mining methods are useful in discovering obscure patterns in agricultural data. These insights help farmers in making wise decisions regarding risk management, timely interventions, and optimization of the yields.

3.2 Data Mining:

Data Mining: This process is used to extract, alter, aggregate, and predict useful information held in the giant data so that it can be eliminated from some of the models and further transformed into usable information that can be utilized on different occasions. [4]. K. Lata, B. Chaudhri, Expectation of crop yield using data mining techniques.

3.3 Machine Learning:

Machine learning (ML) is capable of manipulating designs, associations, and searches of information in datasets. They ought to be ready with the assistance of datasets, and the final findings ought to be considered with reference to the historical experience. The prescient model is developed with reference to several highlights, and therefore, model limits are sorted out within the confirmable data during the exercise phase. During the testing phase, some of the verifiable data that were not previously available are incorporated to evaluate model performance [5]. A. Kassahun and T. van Klompenburg, “Harvest yield prediction using AI: An organized review,” Computers and Electronics in Agriculture, vol. 177, Oct. 2020.

3.4 Supervised Learning:

It is an educative technique that acknowledges input that occurred to yield what is charted out in commitment to yield. Should unattended inclining occur, however, we would know nothing of the specified yield in this acquiring, for which we should be conditioning the model so as to acquire the desired yield. [2]. Application and construction of the collect yield figure model in agribusiness, Global Diary of Logical and 2020 January, Research Volume 8, Issue 1.

3.5Decision Tree

A decision tree is an algorithm of supervised machine learning, which applies a hierarchical, tree-based framework of decision-making and prediction. It can be considered a flowchart, and every inside node resembles a test on an input feature, every branch resembles the finding of the test, and every leaf (terminal) node contains a label of a class or a continuous value of the target variable [2]. The algorithm is configured to project the input features onto a given output because it learns simple decision rules that are induced by the data.

3.6SVR:

Provision Vector Regression (SVR) The concept of Provision Vector Regression is comparable to that of linear regression, where the objective is to estimate a continuous output based on a linear expression: w characterizes the weight and b the bias. This linear form is assumed to be a hyperplane in a higher-dimensional space in SVR.

i. Support Vectors: These are the points nearest to the hyperplane on either side. These are important since they determine the margin and the location of the regression line.

ii. Goal: The goal of SVR is to approximate a line (hyperplane) within a given tolerance.

iii. Based on the observed data, reducing prediction error without considering the points in the margin. In contrast to other common regression models that aim at minimizing the change between the forecast and the real prices, SVR is aimed at molding the model to make most of the points within the defined margin.

3.7 Multinomial NB:

Multinomial Naive Bayes is a probabilistic machine learning algorithm that is typically applied when the task at hand involves text organization, like junk detection, sentiment analysis, or document classification. Main points: According to Bayes' theorem, it calculates the likelihood of any one of the classes (or labels), given the input features. Assumes that features (such as word counts or term frequencies) are conditionally independent based on the class. Multinomial variant: The specific one is well-suited when dealing with count-like characteristics, including the word frequency in a document. The model is effective and efficient on large-scale text classification tasks due to the prediction of the most probable label.

Equation 1: Multinomial NB

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

We are working out the likelihood of class A when indicator B is given.
P(B) = How common class A is overall
P(A) = How often feature B appears in class A
We employed the probability of class A when pointer B is given.
P(B) = How often feature B appears in class A
P(A) How common is class A overall

3.8 Linear Regression:

Linear regression is an arithmetical tool that is employed to estimate the association between dual variables. According to the equation, YYY (the output) is determined as the dependent variable, and XXX (the input) is the independent variable.

Equation 2: Linear Regression

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\varepsilon$$

This calculation serves as the foundation for forecasting upcoming crop yields using historical data, supporting informed supervision and the optimization of agricultural efficiency.

3.9 Gradient Boosting Regressor (GB):

Gradient Boosting Regressor (GB): This is a regression model employed for predicting continuous dependent variables. <|human|>. This is a type of regression used in predicting continuous dependent variables. Gradient Boosting Regressor (GB): is a regressor model that is used when there are no explanatory variables in the data. Gradient boosting Regression divides the boundary between the constant supposition and the identified true value. This has been termed as an extra qualification. Then the inclination that facilitates backsliding is not a good one that connects features with that additional.

3.10 XGB Regressor:

XGBoost is also known as the Risky Gradient Boosting; along with that, it is an application of slant-helping-trees estimation. XGBoost is a famous synchronized simulated intelligence framework, which possesses the features of speed of estimation, parallelism, and execution.

3.11 SGD Regressor:

It (Stochastic Gradient Descent) is actually a simple, yet highly successful method of addressing the problem of training straight classifiers and regressors in bent misfortune works, such as (straight) Support Vector Machines and Strategic Regression. In spite of the long history of SGD diffusion among the AI people group, it has gained significant attention recently as far as learning on a large scale is concerned.

3.12 Kernel Ridge:

Part stunt Edge relapse (L2-regularization and straight least squares) is combined with the part stunt to produce kernel ridge relapse (Kernel ridge relapse). Consequently, it develops a speedy acquisition ability in the discipline that is facilitated by the varied pieces of information. This differs in the non-direct capacity of the non-straight in-line sections in the main space. Elastic Net Regression: It is the quantities to which the variables will apply; however, it does not mean that all the irrelevant coefficients will be eliminated, and this is one of the disadvantages of the ridge regression compared to the Elastic Net Regression (ENR).

Equation 3: Elastic Net Regression

$$\hat{\beta} = \underset{\beta}{\boldsymbol{argmin}} \quad \{\sum_{i=1}^{n} \frac{1}{2n} (y_i - X_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \}$$

3.13 Bayesian Ridge Regression:

The Bayesian regression enables the intrusion of an imperfect system of characterization into a sufficient amount of information or vaguely coded information through the merging of immediate relapse with the aid of likelihood merchants versus the point measures. The response y is believed to be the circulation of likelihood rather than viewing it as an individual amount. The solution y is presented as a number and circulated about the Xw aw in an attempt to obtain a complete specimen of probabilistic circulation.

Equation 4: Elastic Net Regression.

$$P(y \mid \mathrm{X}, \mathrm{w}, \alpha) = \mathrm{N}(\mathrm{y} \mid \mathrm{Xw}, \alpha)$$

4. Training Data and Testing Data

Training data is the set of information that is used to identify predictive relationships among the input variables, as well as output values. The data is used in the training of models that are used in machine learning, artificial intelligence, statistical analysis, and genetic programming. In the course of training, the designs and relations that exist in the data are learned by the model. Testing data, in turn, is an independent part of the dataset, which is not utilized in training. It is given to the trained model to test its performance and generalization capability on unknown data. This is done to ascertain the extent to which the created model is acceptable to predict what truly happens in the field.

4.1 Data Set:

In the majority of research, agricultural data is used in crop yield prediction; this is in the form of CSV (.csv) files that are widely applied by researchers and agricultural professionals.

The model is trained through supervised learning, where historical data of known values of the output is introduced.

The data has several attributes, among them:

- Region Name State
- Soil Type / Tenacity
- Temperature

These features collectively help in estimating crop yield accurately.

4.2 Investigational Result and Comparison

This section involves completing the endorsement of a data set with the help of different calculations; a juxtaposition of experimental outcomes is made and presented in a fragment.

4.3 Confusion Matrix of Linear Regression:

We employed a linear regression model on the dataset, resulting in an overall predictive accuracy of 87%, which demonstrates the model's effectiveness in capturing underlying patterns within the data and its potential suitability for reliable analytical and decision-support applications.

4.4 Confusion Matrix of Multinomial NB:

We implemented a Multinomial Naïve Bayes classifier on the dataset, achieving an accuracy of 87%, which indicates its strong capability to effectively model the probabilistic relationships among features and deliver reliable classification performance.

Table 1. Comparison of Techniques

Model	Accuracy	Macro Precision	Macro Recall	Macro F1-Score	Weighted Precision	Weighted Recall	Weighted F1-Score
Linear Regression	0.89	0.80	0.78	0.79	0.89	0.89	0.89
Multinomial Naïve Bayes	0.90	0.81	0.82	0.81	0.91	0.90	0.90
XGB Regressor	**0.99**	**0.99**	**0.98**	**0.99**	**0.99**	**0.99**	**0.99**
Decision Tree	0.95	0.90	0.92	0.91	0.95	0.95	0.95
Kernel Ridge	0.90	0.83	0.78	0.80	0.90	0.90	0.90
Bayesian Ridge	0.89	0.81	0.79	0.80	0.89	0.89	0.89
Elastic Net	0.84	0.79	0.78	0.78	0.83	0.84	0.83
Gradient Boosting Regressor	0.93	0.92	0.86	0.88	0.93	0.93	0.93
SVR	0.87	0.79	0.78	0.79	0.87	0.87	0.87

Table 2. Comparison of Techniques

Model	MAE	MSE	RMSE	Accuracy (%)
SVR	15.25	11.34	3.90	84.75
Linear Regression	12.56	8.65	3.54	87.44
GB Regressor	8.07	4.16	2.84	91.93
XGB Regressor	8.52	4.61	2.92	91.48
Decision Tree Regressor	6.73	2.82	2.59	93.27
Kernel Ridge	12.11	8.20	3.48	87.89
Elastic Net	15.25	11.34	3.90	84.75
Bayesian Ridge	13.45	9.54	3.67	86.55
Multinomial Naive Bayes	13.45	9.54	3.67	86.55

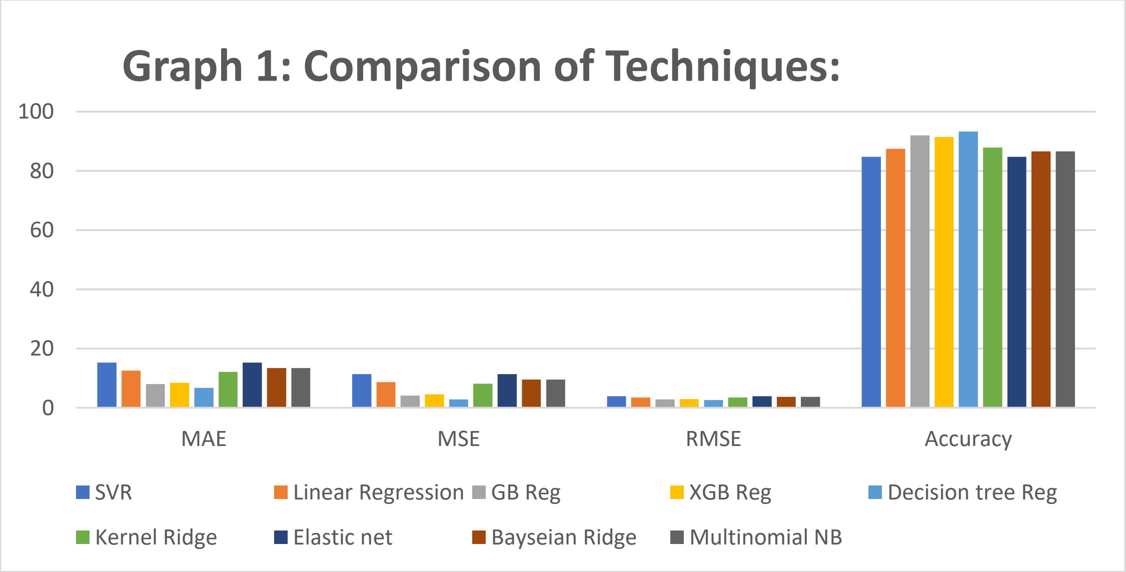

FIGURE 1: Comparison of Techniques Graph

5. Conclusion:

This study undertakes a comprehensive and critical examination of data mining and machine learning techniques to evaluate their effectiveness in accurately predicting crop yield. The results indicate that smart predictive models, especially the decision trees, may greatly improve the precision and accuracy of yield forecasts. The proposed framework can offer effective information to farmers and the policymaking body by considering the environmental and agronomic factors, which can aid in better agricultural planning, risk management, and food security. The paper demonstrates the prospects of data-driven agriculture in Pakistan and other developing economies, in which precise yield forecasting can lead to the establishment of sustainable food production and financial stability.

5.1 Future work:

Future studies can also be used to enhance the accuracy of prediction by adding more attributes like soil nutrient contents (e.g., nitrogen levels), humidity, evapotranspiration, fertilizer application rates, remote sensing, and satellite imagery data. Real-time data sources and deep learning methods can also be integrated to potentially improve system adaptability and scalability to make more accurate and region-specific decision support in agriculture.

5.2 Adding More Attributes for Better Predictions:

The Additional Attributes Improved the Guesses: The future work will be structured according to the analysis of the entire system statistics and will be dedicated to the realistic systems employed on the efficiency of the offered calculation. Such a kind of way of dealing with anticipation is not restricted to agribusiness. The compilation and reconstruction is a skilled device in the information mining market that may be utilized in a wide range of applications. The proposed system can include several properties that can better predict the yield in the future, and they include

i. Soil type.
ii. Nitrogen amount of the soil.
iii. The intermittent, non-occasional yields.
iv. Utilization of fertilizers.
v. Humidity.

6. References

K. Lata and B. Chaudhari, “Crop yield prediction using data mining techniques and machine learning models for a decision support system,” *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 6, no. 4, pp. 1–6, Apr. 2019. [Online]. Available: http://www.jetir.org. Accessed: Jun. 1, 2022.

S. G., “Design and implementation of crop yield prediction model in agriculture,” *International Journal of Scientific & Technology Research (IJSTR)*, vol. 8, no. 1, pp. 1–6, 2020. [Online]. Available: http://www.ijstr.org. Accessed: Jul. 19, 2022.

S. K., S. Barker, and S. Kulkarni, “Analysis of crop yield prediction using data mining technique,” *International Journal of Research in Engineering and Technology (IJRET)*, vol. 7, no. 5, pp. 1–5, 2020. [Online]. Available: http://www.irjet.net. Accessed: May 10, 2022.

D. Ramesh and B. Vardhan, “Analysis of crop yield prediction using data mining techniques,” *International Journal of Research in Engineering and Technology (IJRET)*, vol. 4, no. 1, pp. 1–4, 2022. [Online]. Available: http://www.ijret.org. Accessed: Jul. 6, 2022.

T. van Klompenburg, A. Kassahun, and C. Catal, “Crop yield prediction using machine learning: A systematic literature review,” *Computers and Electronics in Agriculture*, vol. 177, pp. 1–18, 2020.

I. Aroquiaraj and P. Surya, “Crop yield prediction in agriculture using data mining predictive analytic techniques,” *International Journal of Research and Analytical Reviews (IJRAR)*, vol. 5, no. 4, pp. 1–5, 2018. [Online]. Available: http://www.ijrar.org. Accessed: Jun. 16, 2022.

N. T., "Prototyping model," *Binary Terms*, 2020. [Online]. Available: https://binaryterms.com/prototyping-model.html. Accessed: Apr. 6, 2022.

"UML – Activity diagrams," *Tutorials Point*, 2020. [Online]. Available: http://www.tutorialspoint.com. Accessed: Jun. 7, 2022.

R. A., M. Vijay, and S. Rajpurohit, "Crop yield prediction data set," *Kaggle*, 2019. [Online]. Available: http://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset. Accessed: Mar. 4, 2022.

D. Haak, "Crop yield: Definition & consequences," *Study.com*, 2022. [Online]. Available: https://study.com/academy/lesson/crop-yield-definition-consequences.html. Accessed: Jun. 15, 2022.

"Non-functional requirements: Examples, types, how to approach," *Alex Soft*, 2022. [Online]. Available: http://www.alexsoft.com/blog/non-functional-requirements. Accessed: Jul. 11, 2022.

"Crop yield gap analysis Pakistan | 2020," *Zarai Taraqiati Bank Limited (ZTBL)*, 2020. [Online]. Available: https://ztbl.com.pk/media-enter/research-publications/. Accessed: Jun. 10, 2022.

S. Bhojani and N. Bhatt, "Review of literature of data mining techniques for crop yield prediction," *International Journal of Scientific Research*, vol. 6, no. 12, pp. 455–457, 2017. [Online]. Available: https://www.researchgate.net/publication/326439354. Accessed: Jul. 6, 2022.

"Confusion matrix," *Engati*, 2022. [Online]. Available: https://www.engati.com/glossary/confusion-matrix. Accessed: Jul. 19, 2022.

S. Patil, R. M., P. K., and V. H., "Crop yield prediction using machine learning," *International Journal of Engineering Science and Computing (IJESC)*, vol. 10, no. 7, pp. 1–2, 2020. [Online]. Available: http://www.ijesc.org. Accessed: Apr. 9, 2022.