Crop Yield Prediction and Classification using the Agriculture Resources and Data Mining Techniques

Pramod Kumar Dwivedi[1]*, Dr. Prabhat Pandey[2]

[1] Research Scholar, Awadhesh Pratap Singh University, Rewa, Madhya Pradesh, India

Email: ashwanikhajuraho@gmail.com

[2] Professor, Department of Computer Science, Awadhesh Pratap Singh University, Rewa, Madhya Pradesh, India

Abstract - This study investigates the application of data mining techniques in meteorological forecasting to enhance crop yield prediction accuracy Moreover, the integration of meteorological forecasting with agronomic models and geographical information systems (GIS) facilitates site-specific crop management and precision agriculture practices. By combining meteorological data with soil properties, crop phenology, and socio-economic factors, data mining techniques enable the development of predictive models that enhance crop yield potential and optimize resource allocation. By leveraging advanced data analytics and machine learning algorithms, agricultural stakeholders can make informed decisions, mitigate risks, and enhance productivity in the face of changing climatic conditions and environmental uncertainties.

Keywords: Agriculture, Resources, Crop Yield Prediction and Data Mining, Techniques

---------------------------------x---------------------------------

INTRODUCTION

The Indian economy and Indian culture have long relied on agriculture. Involvement of the country's biggest population, whether via direct or indirect work, self-employment, or partial employment, results in a substantial contribution to GDP. Modern society owes its very existence to agriculture. Agriculture, the practice of cultivating domesticated species for the purpose of producing food surpluses, was a critical innovation in the emergence of sedentary human civilization. The general definition of agriculture is the study of farming and other land-based pursuits. The practice of raising various living things for human subsistence and improvement is known as agriculture, farming, or husbandry. This includes not just animals but also plants and fungus. Plantations and crops are the principal products of cultivation. Animals are raised for meat, wool, and other goods. Aqua-products, such as fish, are also produced by agriculture. Agriculture is all about us: in the food we eat, the clothes we wear, the timber and materials needed to build houses, and in the gardens, flowers, and greenery that surround us. Modern farming encompasses a wide range of activities, such as managing inputs, implementing good agricultural practices, protecting plants, managing plant nutrients, managing irrigation, harvesting crops, processing and adding value to farm products, marketing and distributing these products, dealing with agri-inputs supply and service industries, addressing issues related to human health, nutrition, and food consumption, practicing wise land and water resource conservation, and addressing related economic, social, political, environmental, and cultural concerns.

Agricultural science is the academic discipline that focuses on farming. Agricultural practices have been around for a long time, influenced by many various civilizations, climates, and technological advancements. All farming, however, depends on methods to increase and preserve arable land appropriate for domesticated animal husbandry. Despite rising interest in sustainable farming practices like permaculture and organic farming, industrial agriculture based on large-scale monoculture is already the standard in industrialised nations. The growth and preservation of various crops depend on each and every one of the many cultivating procedures. Mechanical agriculture has become the dominant method of modern farming due to its emphasis on large-scale monoculture. Nevertheless, widespread agricultural mechanization is very difficult to implement in poorer nations due to the tiny and scattered land holdings. However, with an eye towards heightened concentration on profit margins and market demands, monoculture is rapidly becoming the norm. When it comes to a wide variety of crops, including wheat, cotton, peanuts, castor, rice, milk, mangoes, and many more, India is among the top three global producers. The world's largest herds of buffalo and domestic cattle are located in India. An important issue in India is the efficient use of agricultural resources, such as seeds, water, composts, fertilizers, and pesticides. Due to factors such as better soil quality, more arable land, improved agriculture techniques, higher yields from livestock and fisheries, more efficient use of water, and so on, farming in India has been better in recent decades. Additional assessment, inspection, and optimal designation of these assets with micro farming circumstances in mind would allow for the practical improvement of farming in India, which in turn will enhance agricultural production in the country.

LITERATURE REVIEW

“Mishra (2018)” Using data mining methods on an agricultural dataset, this research also focuses on implementing a system to estimate crop yields. Using the WEKA tool, the first stage compared the performance of several classifiers, including J48, LWL, LAD Tree, and IBK. Other metrics used to compare classifiers include root-mean-square error, mean absolute error, and relative absolute error. Based on the findings of this investigation, IBK obtains the best accuracy and LAD Tree has the worst.

Balducci, Impedovo, & Pirlo, (2018) Crop production forecasting and comparing different machine learning algorithms are two examples of the kinds of practical tasks that this project aims to build and execute. This research’s findings reveal that neural network models can achieve total crop forecasts for apples and pears on the Istat dataset with success rates close to 90%. On the other hand, when it comes to the second task, it becomes clear that, given the type and nature of the dataset, CNR scientific data is best served by regression models and polynomial predictive.

The paper submitted by H. Kerdiles, et al. (2017) concentrated on the use of Crop Statistics Tool (CST), a free, stand-alone programme for calculating and forecasting crop yields. It is safe to say that the CST's findings are reliable, accurate, and very effective. The author recommended enhancing CST's graphics and, in terms of statistical approaches, revising scenario analysis to take into consideration the indicators' temporal profiles rather than their value at a certain instant.

Gandhi, N., et al., (2016) displayed the outcomes acquired from the dataset consisting of 27 districts in the Indian state of Maharashtra using the WEKA tool and the unique machine learning approach known as SMO classifier. Using the experimental data, they determined that compared to the SMO classifier, which had the lowest accuracy and produced an unspecific result, other classifiers including Naïve Bayes, BayesNet, and Multilayer Perceptron attained the best levels of accuracy, sensitivity, and specificity. They came to the conclusion that the other classifiers would work better with the existing data set.

Gandhi, N., and Armstrong, J. L., (2016) drawn attention to the importance of data visualisation in gaining a broad understanding of the ways in which different variables impact agricultural productivity. In order to establish a connection between the climatic element and agricultural production, they investigated many data visualisation methods. Their experimental results demonstrate that the J48 and LADtree algorithms provide the most precise and precise results, whereas LWL yields the most inconsistent and general results with the worst precision

RESEARCH METHODOLOGY

The practice of forecasting future crop yields and producing goods in anticipation of those yields is known as crop prediction. Historically, farmers would use their expertise in a particular area and crop to make predictions about when crops will be ready to harvest. The analysed datasets may be predicted by this system, which employs data processing techniques. The expected kind may determine crop yields. On top of that, let's assume this dataset already has dimensionality d, where $d < DI, \mathit{and\ frequently}\ d <<< DI$, Currently, according to scientific standards, the points in dataset E are either lying on or very close to a dimensionality d variable, which indicates they are embedded in the DI dimension.

By absorbing the data's geometry to the maximum extent feasible, dataset dimensionality reduction techniques transform dataset E, which has dimensionality DI, into a new dataset F, which has dimension d. Curve fitting, sometimes called regression analysis, finds the "best fit" line or curve for a set of data points. This section makes use of Matlab's curve fitting function to determine and examine the relationship between rainfall and agricultural production. Cluster analysis, also known as clustering, is the process of organising data into sets wherein items within each set are more similar to one another than those outside of it. To estimate the support vector machine (SVM) that would have been requested from the issue, a tolerance margin (epsilon) is specified while regression is being considered. This is true, but there is another, more nuanced factor to think about, and the algorithm is more involved as a result.

RESULT

CROP YIELD PREDICTION

Data mining methods are applied to the given data set, and the findings are discussed in this chapter. We have also spoken about how to use the data to uncover new information. Additionally, the chapter concludes by discussing ways in which the suggested work may be improved.

Results Obtained

The outcomes and step-by-step screenshots from each stage of the suggested model down below.

The initial user interface (UI) of our suggested model, which accepts the training and test datasets as inputs, is shown in Figure 1. The user interface (Figure 2) after data selection for training and testing.

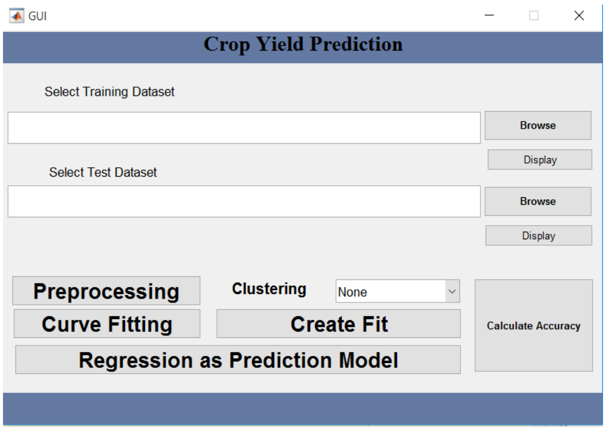
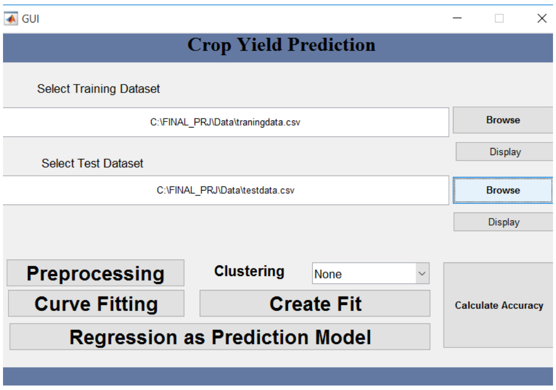


Figure 1 Main UI of Proposed Model


C:\FINAL_PRJ\Data\traningdata.csv

C:\FINAL_PRJ\Data\testdata.csv

Figure 2 UI of Model after selecting Training and Testing Data

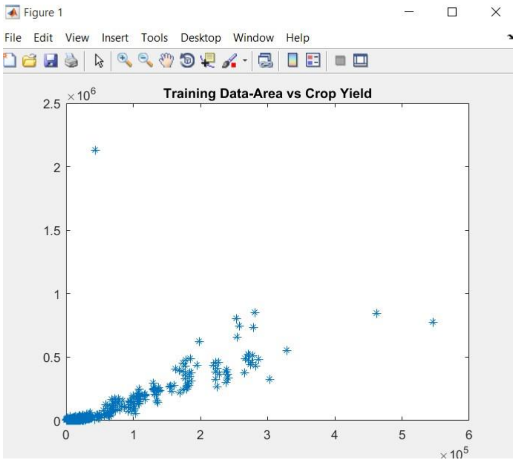


Figure 3 2D Plot of selected Training Data: Area vs Crop Yield

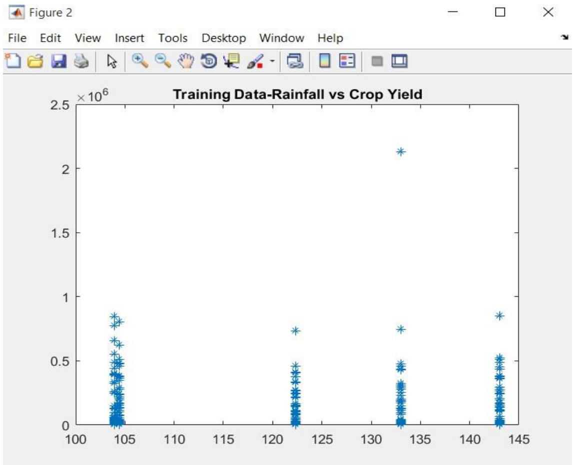


Figure 4 2D Plot of selected Training Data: Rainfall vs Crop Yield

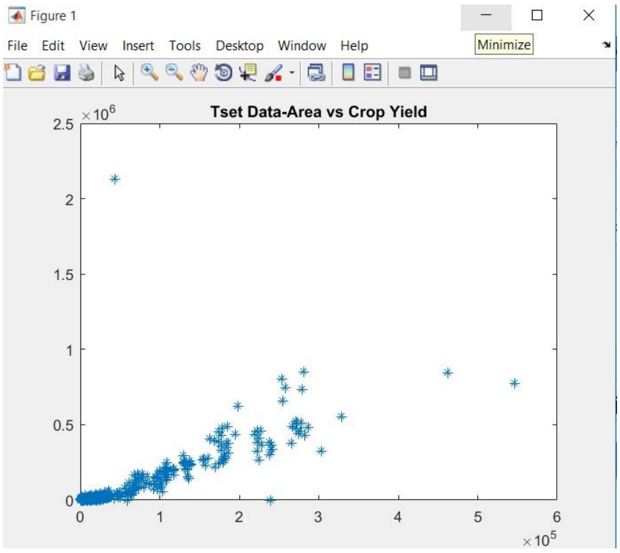


Figure 5 2D Plot of selected Testing Data: Area vs Crop Yield

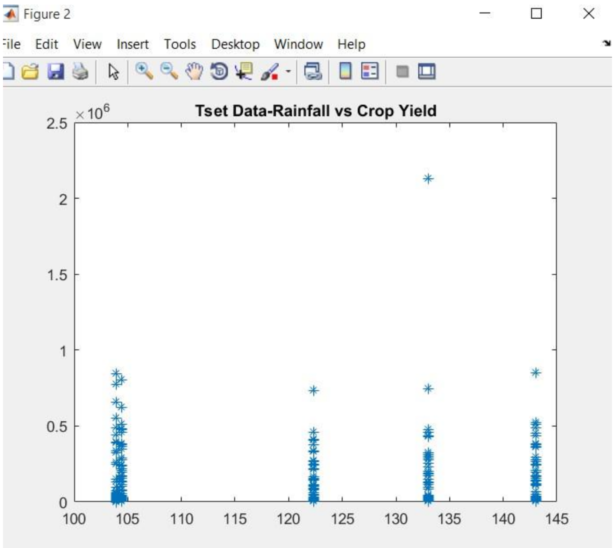


Figure 6 2D Plot of selected Testing Data: Rainfall vs Crop Yield

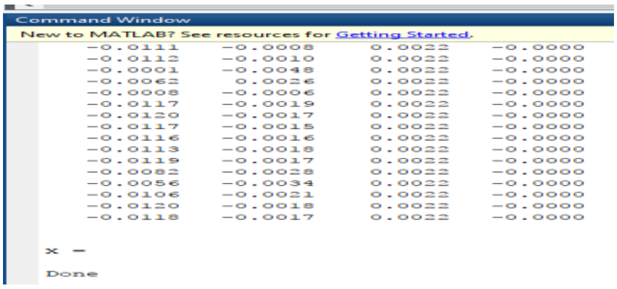
Pre-processing:

The suggested model employs principal component analysis (PCA) to reduce dimensions, as shown in Figure 7, which displays the PCA coefficients.

The PCA score matrix is determined as: Matrix of covariance

$$Y = W'X \dots (1)$$

Where X is the matrix of input data and W' is the matrix of weights.

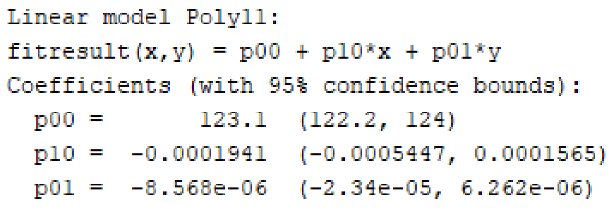


```
Command Window
New to MATLAB? See resources for Getting Started.
   -0.0111   -0.0008    0.0022   -0.0000
   -0.0112   -0.0010    0.0022   -0.0000
   -0.0001   -0.0048    0.0022   -0.0000
   -0.0062    0.0026    0.0022   -0.0000
   -0.0008   -0.0006    0.0022   -0.0000
   -0.0117   -0.0019    0.0022   -0.0000
   -0.0120   -0.0017    0.0022   -0.0000
   -0.0117   -0.0015    0.0022   -0.0000
   -0.0116   -0.0016    0.0022   -0.0000
   -0.0113   -0.0018    0.0022   -0.0000
   -0.0119   -0.0017    0.0022   -0.0000
   -0.0082   -0.0028    0.0022   -0.0000
   -0.0056   -0.0034    0.0022   -0.0000
   -0.0106   -0.0021    0.0022   -0.0000
   -0.0120   -0.0018    0.0022   -0.0000
   -0.0118   -0.0017    0.0022   -0.0000

x =

Done
```

Figure 7 PCA Coefficients

We have eliminated outliers and decreased the dataset's dimensions using PCA coefficients. To determine the optimal fit between the independent and dependent variables, additional datasets are provided to Matlab's cftool.

Figure 8 shows the best-fit equation below.



```
Linear model Poly11:
fitresult(x,y) = p00 + p10*x + p01*y
Coefficients (with 95% confidence bounds):
  p00 =       123.1  (122.2, 124)
  p10 =  -0.0001941  (-0.0005447, 0.0001565)
  p01 =  -8.568e-06  (-2.34e-05, 6.262e-06)
```

Figure 8 Curve fitting Result (Best fit)

Figures 9 and 10 show the input dataset clustering. We found that the number of outliers is lower while using k-medoid clustering.

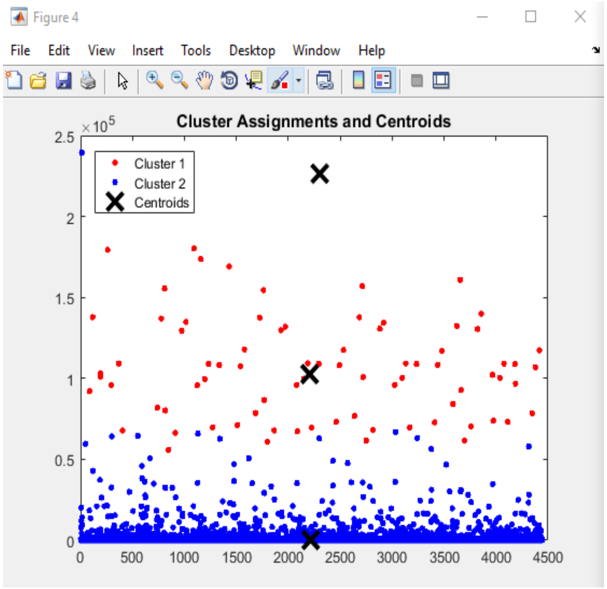


Figure 9 Data Clustering in K-means

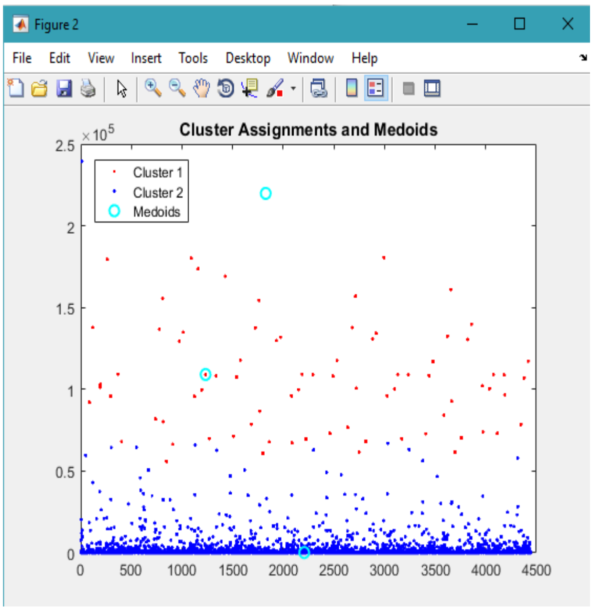


Figure 10 Data Clustering in K-medoid

The accuracy of the proposed machine learning model may be evaluated by referring to Figure 12, which displays the forecast value of crop yield, and Figure 11, which shows the 2D plot of actual values versus anticipated values.

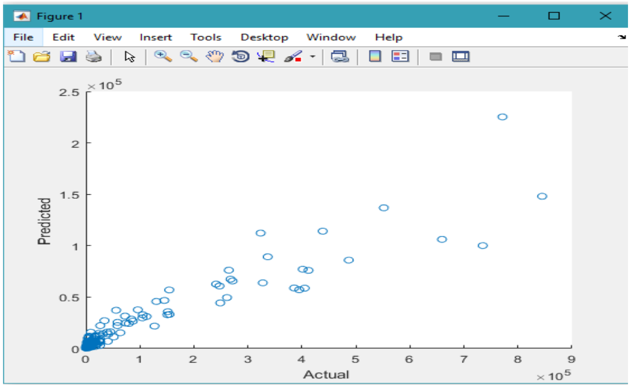
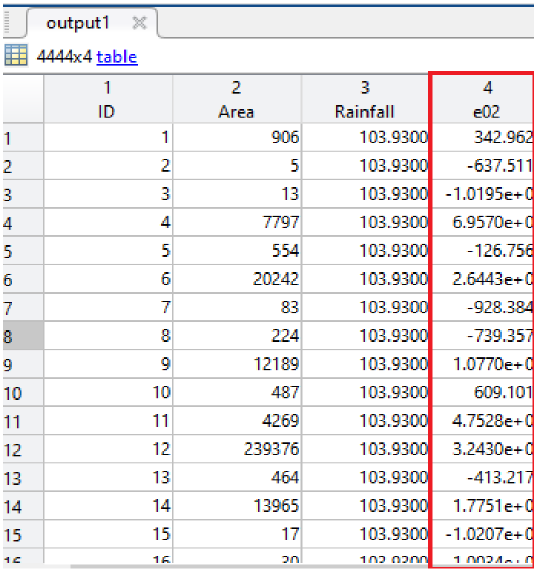


Figure 11 2D Plot of actual; values vs Predicted Values



output1

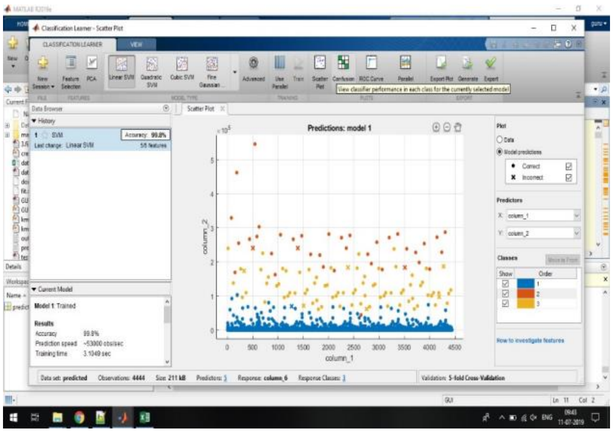
4444x4 table

	1 ID	2 Area	3 Rainfall	4 e02
1	1	906	103.9300	342.962
2	2	5	103.9300	-637.511
3	3	13	103.9300	-1.0195e+0
4	4	7797	103.9300	6.9570e+0
5	5	554	103.9300	-126.756
6	6	20242	103.9300	2.6443e+0
7	7	83	103.9300	-928.384
8	8	224	103.9300	-739.357
9	9	12189	103.9300	1.0770e+0
10	10	487	103.9300	609.101
11	11	4269	103.9300	4.7528e+0
12	12	239376	103.9300	3.2430e+0
13	13	464	103.9300	-413.217
14	14	13965	103.9300	1.7751e+0
15	15	17	103.9300	-1.0207e+0
16	16	20	103.9300	1.0034e+0

Figure 12 Predicted Values

Using the Matlab classification learner tool, we conducted an experiment to evaluate the suggested prediction model by passing the predicted data to a quadratic SVM classifier.

For k-medoid clustering, the accuracy of the suggested model is shown in Figure 13 when the predicted data is sent to the SVM classifier.



View classifier performance in each class for the currently selected model

1 SVM Accuracy: 99.8%
Last change: Linear SVM 5/5 features

Current Model

Model 1: Trained

Results
Accuracy 99.8%
Prediction speed ~53000 obs/sec
Training time 3.1049 sec

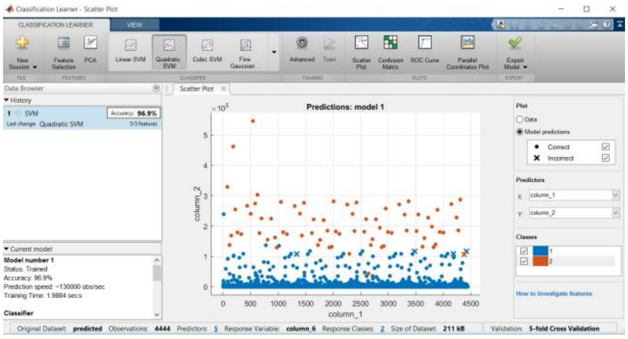
X: column_1
Y: column_2

Show	Order
☑	🟦 1
☑	🟥 2
☑	🟨 3

Data set: predicted Observations: 4444 Size: 211 kB Predictors: 5 Response: column_6 Response Classes: 3 Validation: 5-fold Cross-Validation

Figure 13 Accuracy of the proposed model in case of k-medoid clustering

In the instance of k-means clustering, Figure 14 displays the accuracy of the suggested model when the forecasted data is provided to the SVM classifier.



1 SVM Accuracy: 96.9%
Last change: Quadratic SVM 5/5 features

Current model

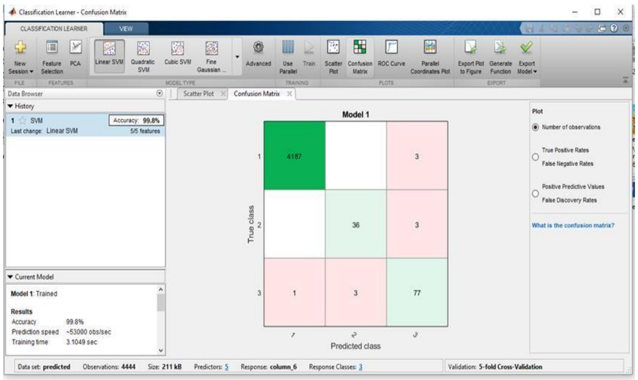
Model number 1
Status: Trained
Accuracy: 96.9%
Prediction speed: ~130000 obs/sec
Training Time: 1.9884 secs

Classifier

X: column_1
Y: column_2

Original Dataset: predicted Observations: 4444 Predictors: 5 Response Variable: column_6 Response Classes: 2 Size of Dataset: 211 kB Validation: 5-fold Cross Validation

Figure 14 Accuracy of the proposed model in case of k-means clustering



1 SVM Accuracy: 99.8%
Last change: Linear SVM 5/5 features

True class \ Predicted class	1	2	3
1	4187		3
2		36	3
3	1	3	77

Current Model

Model 1: Trained

Results
Accuracy 99.8%
Prediction speed ~53000 obs/sec
Training time 3.1049 sec

Data set: predicted Observations: 4444 Size: 211 kB Predictors: 5 Response: column_6 Response Classes: 3 Validation: 5-fold Cross-Validation

Figure 14 ROC curve

From what we can see in Figures 13 and 14, it is clear that clustering with k medoid yields better results than k means.

Accuracy predictions are shown in Figure 14, which is a ROC curve. The curve includes True Positive and False Negative values.

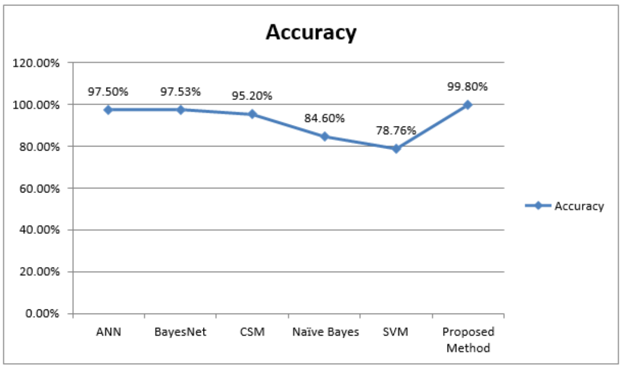


Figure 15 Comparison of accuracy with previous work

The average accuracy is growing as the testing dataset becomes larger, as shown in Figure 16.

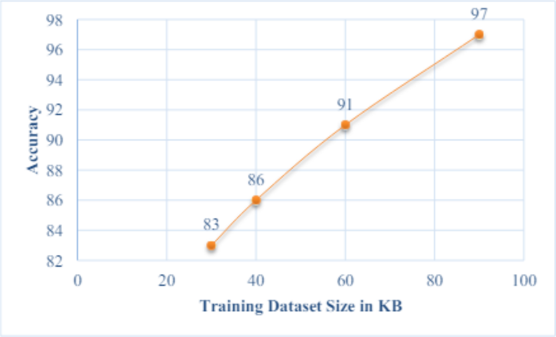


Figure 16 Accuracy of the proposed model

CONCLUSIONS

To determine the relationship between the dependent and independent variables in the proposed model, we used the curve fitting tool in MATLAB. When it came time to apply a machine learning classifier, we went with support vector regression and ended up with an average accuracy of 97%. The major goal of this project is to improve the precision of agricultural yield forecast. Research in this field is still in its early stages, but it holds great promise for data miners looking to have a positive impact on society. A huge obstacle to ongoing agricultural growth is the weight that comes from recurring asset constraints, increasing possession fragmentation, frequent climate change, growing input expenditures, and post-harvest disasters. Farming and agriculture provide a large portion of India's economy and employ a large portion of the people. The government reports that the Monsoon rains are responsible for 60% of the country's agricultural output. Consequently, developing a prediction model for agricultural production prediction is necessary, and understanding the elements impacting crop yield is of utmost importance.

REFERENCES

1. Mishra, S. P. (2018). Use of Data Mining in Crop Yield Prediction. Proceedings of the Second International Conference on Inventive Systems and Contro (pp. 796-802). IEEE.

2. Balducci, F., Impedovo, D., & Pirlo, G. (2018). Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement. Machines, 6, 38.

3. Gandhi, N., Armstrong J. Leisa, (2016), Rice crop yield prediction using artificial neural networks, IEEE International Conference on Technological Innovations in ICT For Agriculture and Rural Development (TIAR) 105.

4. Gandhi, N., Armstrong J. Leisa, Owaiz P., (2016), Predictingrice crop yield using bayesian networks, Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, Jaipur, India.

5. Anusha A. Shettar, S. A. (2016). Efficient data mining algorithms for agriculture data. International Journal of Recent Trends in Engineering & Research, 142-149.

6. Rajshekhar Borate., “Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in India”, International Journal of Novel Research in Computer Science and Software Engineering,Vol. 3, Issue 1, pp: (34-37), April 2016.

7. D Ramesh, B Vishnu Vardhan, “Analysis of Crop Yield Prediction using Data Mining Techniques”, International Journal of Research in Engineering and Technology (IJRET),Vol.4, 2015.

8. Veenadhari, S., Bharat Misra, D Singh, “Data mining Techniques for Predicting Crop Productivity – A review article”, IJCST, International Journal of Computer Science and technology march 2011.

9. DakshayiniPatil, Dr. M .S Shirdhonkar, “Rice Crop Yield Prediction using Data Mining Techniques: An Overview”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 5, ISSN: 2277 128X,2017.

10. Ramesh A. Medar and Vijay. S. Rajpurohit “A Survey of data mining techniques for crop yield prediction”, IJARCSMS, Volume 2, Issue 9, September 2014 pg. 59-64.

11. D, Ashok Kumar and Kannathasan, N (2011) A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining, International Journal of Computer Science Issues, Vol. (8).

12. Geetha, M.C. (2018), A Survey and Analysis on Regression Data Mining Techniques in Agriculture, International Journal of Pure and Applied Mathematics, Vol 118 No. 8, 341-347.

13. Ramesh Vamanan, K. Ramar (2011), Classification of agricultural land soils a data mining approach, International Journal on Computer Science and Engineering, Vol. 3(1).

14. Vagh, Y. (2012), An investigation into the effect of stochastic annual rainfall on crop yields in South Western Australia, Paper presented at the International Conference on Knowledge Discovery.

15. Dissanayake, D. & Rathnayake, R. & Chathuranga, Gihan. (2023). Crop Yield Forecasting using Machine Learning

Techniques - A Systematic Literature Review. KDU Journal of Multidisciplinary Studies. 5. 54-65. 10.4038/kjms.v5i1.62.

Corresponding Author

Pramod Kumar Dwivedi*

Research Scholar, Awadhesh Pratap Singh University,
Rewa, Madhya Pradesh, India

Email: ashwanikhajuraho@gmail.com