An Improved Decision Tree Algorithm in Machine Learning Concept for Crop Yield Prediction

Dr. S Jeyalaksshmi

Associate Professor, Department of BCA and IT,
Vels Institute of Science Technology & Advanced Studies
Email : jlakshmi.scs@velsuniv.ac.in

ABSTRACT:

Farming is very important to the Indian culture. Technological progress has opened the door to better farming conditions and more effective agricultural policymaking. The agricultural database is accessed using a Python interface. Data miners using Jupyter Notebook to predict crop output. Weather, temperature, reference crop, evapotranspiration, area, production, and yield statistics are all part of the data collection, which covers the years 2000-2018. Support vector machines, K-Nearest Neighbors, K-Means clustering, and the Bayesian network approach are data mining algorithms that provide pinpoint accuracy.

Keyword: *Agriculture, Machine Learning, crop-prediction, Random forest algorithm, KNN, SVM, python*

I. INTRODUCTION

Crop production may be a relatively new development, depending on the input characteristics of the soil and its state. The process elements in agriculture might differ from one farm or farmer to another. Collecting the same data across a somewhat larger region might provide equally disheartening challenges. However, the Indian Meteoric Department tabulates data on the environmental status of the Republic of India for every 1 sq. m of land across the district's multiple portions. It is common practice to use these enormous datasets to forecast the impact of these variables on the main crops cultivated in a certain area. Scientists from all across the globe have developed their own special methods of agricultural and related field prediction. Examples of such studies include the following: International agricultural specialists have confirmed that efforts to maximize the use of pesticides in an effort to increase crop yields have been made. The introduction of very high chemical consumption levels is a direct outcome of driving efforts. According to these studies, there is a correlation between pesticide use and higher yields [1]. One of our most vital partner sectors, agriculture, has been greatly affected by recent developments in detection technology, information science, and machine learning (ML) techniques.

With the rise of ML, big data, and HPC, new avenues for interpreting, measuring, and comprehending data-intensive processes in agricultural operational settings have opened up. Computers may learn new tasks without any intervention from a human programmer; this capability is known as "machine learning" (ML). Machine learning is among the most fascinating subfields in computer science right now. Commonly, people would say that machine learning is the same as artificial intelligence (AI). Finally, an automated system that can mimic human learning by analyzing and predicting patterns in data and future events. The machine learning methods are much improved compared to the older set-rule techniques. Several pieces of data or results may be examined by them with ease and precision. To begin learning, the developer must first enter data into the machine's programming language. Computer algorithms may take a broad variety of inputs into account and provide an analysis of any common situation. With a high level of precision, one may anticipate future agricultural yields, nutritional value, etc. It is possible to classify machine learning algorithms into three main types: supervised, unsupervised, and reinforcement.

II. RELATED WORK

Shruthi G. Sangeeta [1]. Machine learning methods including Decision Trees, Polynomial Regression, and Random Forests are used to evaluate the project's performance. Of the three methods we examined for this model, Random Forest yield prediction performed the best. Some of the methods used to categorize the dataset improvements include decision trees, polynomial regression, and forest at random. That being said, we found the proposed model to be a better predictor of crop yield than the existing one. If we use the aforementioned strategy, maybe our country's agricultural methods will improve. Also, farmers might have access to greater financial resources if it helps them enhance agricultural production and minimize losses. The concept might be improved by collaborating with other organizations that are advancing American agriculture, such those who deal with horticulture and sericulture.

Among the many subjects addressed in the 39 papers compiled by Bhawana Sharma et al. [2] are yield prediction, weed detection, disease identification, soil and water management, and many more aspects of agricultural work. Some approaches do better than others when it comes to accuracy. For anyone interested in using machine learning in the agriculture field, that is encouraging news. It is not made clear in any of these crop management algorithms that figuring out when a crop is ready to harvest is their main focus. As a result, monitoring crop maturity using machine learning might be a viable alternative. This paper demonstrates a proposed methodology that uses image processing and machine learning techniques to determine crop maturity from digital images. Incorporating state-of-the-art deep learning and machine learning techniques is a great way to improve upon existing models.

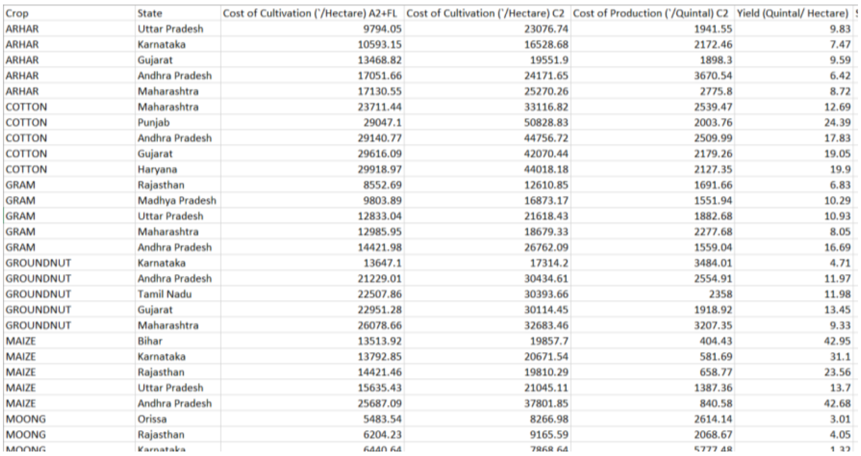
“Dr. Konstantinos G. Liakos and colleagues” [3] By integrating machine learning with sensor data, farm management systems are evolving into full AI systems. This allows them to provide more detailed recommendations for enhancing productivity. There will be a plethora of opportunities for practical and interoperable applications created as a result of the increasing use of ML models in this setting. Current techniques, like those in other domains, ignore the importance of integrating approaches and responses into decision-making in favor of treating them as standalone items. The purpose of so-called "knowledge-based agriculture" is to increase production rates and bio-product quality. Combining automated information recording, data analysis, ML deployment, and decision-making or help will deliver practical advantages that align with this strategy.

III. PROPOSED METHOD

There are three main components to machine learning: supervised learning, reinforcement learning, and unsupervised learning. A kind of machine learning known as "supervised learning" involves the deliberate instruction of the relationship between inputs and outputs. Conversely, in unsupervised learning, the target result is not known before the model is trained. The project's implementation was divided in two.For the fertilizers module, for instance, it could be helpful to predict agricultural output and precipitation.

The database used for this inquiry represents the yields of several crops, including arhar, cotton, lentil, moong, rice, mustard, sugarcane, and wheat, among many others. This includes the following Indian states: Andhra Pradesh, Bhilai, Gujarat, Haryana, Rajasthan, Uttar Pradesh, Orissa, West Bengal, Punjab, Madhya Pradesh, and Maharashtra.

Their supporting price, yields per hectare in quintals, production cost per hectare, and cultivation price per hectare make up the data collected here.



Crop	State	Cost of Cultivation (`/Hectare) A2+FL	Cost of Cultivation (`/Hectare) C2	Cost of Production (`/Quintal) C2	Yield (Quintal/ Hectare)
ARHAR	Uttar Pradesh	9794.05	23076.74	1941.55	9.83
ARHAR	Karnataka	10593.15	16528.68	2172.46	7.47
ARHAR	Gujarat	13468.82	19551.9	1898.3	9.59
ARHAR	Andhra Pradesh	17051.66	24171.65	3670.54	6.42
ARHAR	Maharashtra	17130.55	25270.26	2775.8	8.72
COTTON	Maharashtra	23711.44	33116.82	2539.47	12.69
COTTON	Punjab	29047.1	50828.83	2003.76	24.39
COTTON	Andhra Pradesh	29140.77	44756.72	2509.99	17.83
COTTON	Gujarat	29616.09	42070.44	2179.26	19.05
COTTON	Haryana	29918.97	44018.18	2127.35	19.9
GRAM	Rajasthan	8552.69	12610.85	1691.66	6.83
GRAM	Madhya Pradesh	9803.89	16873.17	1551.94	10.29
GRAM	Uttar Pradesh	12833.04	21618.43	1882.68	10.93
GRAM	Maharashtra	12985.95	18679.33	2277.68	8.05
GRAM	Andhra Pradesh	14421.98	26762.09	1559.04	16.69
GROUNDNUT	Karnataka	13647.1	17314.2	3484.01	4.71
GROUNDNUT	Andhra Pradesh	21229.01	30434.61	2554.91	11.97
GROUNDNUT	Tamil Nadu	22507.86	30393.66	2358	11.98
GROUNDNUT	Gujarat	22951.28	30114.45	1918.92	13.45
GROUNDNUT	Maharashtra	26078.66	32683.46	3207.35	9.33
MAIZE	Bihar	13513.92	19857.7	404.43	42.95
MAIZE	Karnataka	13792.85	20671.54	581.69	31.1
MAIZE	Rajasthan	14421.46	19810.29	658.77	23.56
MAIZE	Uttar Pradesh	15635.43	21045.11	1387.36	13.7
MAIZE	Andhra Pradesh	25687.09	37801.85	840.58	42.68
MOONG	Orissa	5483.54	8266.98	2614.14	3.01
MOONG	Rajasthan	6204.23	9165.59	2068.67	4.05
MOONG	Karnataka	6440.64	7868.64	5777.48	1.32

FIGURE1. Sample dataset

i. Crop Yield Prediction

This module will spit out your predicted crop yield if you input it. Also, you may tell the machine what sort of crop you want by looking at the output. Otherwise, you will get a list of species along with their yields. The steps to execute the algorithm are as follows:

- Step 1: To get crop or yield predictions, for example, just choose the feature you want.
- Step 2: For crop prediction, you'll need to know the soil type and the acreage of the land. We feed these values into the random forest's back-end implementation and get back our predictions. The algorithm provides a response that details the crops together with their projected yields.
- Step 3 : Selecting yield prediction requires the user to provide crop information, soil type, and land area. By inputting these values into the backend's random forest implementation, we may anticipate the harvest's yield. The method's result is the predicted harvest for the given crop.

ii. Random Forest

Like other ensemble classifiers, Random Forests employ a large number of decision tree models to create predictions. An whole fresh subset of the data is randomly selected and used to train each tree. The idea behind random forests is to train the trees using randomly selected subsets of the trees already present in the forest. Instances of classification and regression problems are two examples of potential uses. To find out which class each tree is a part of, we take the average of their votes; to do regression, we use the same approach. What follows is a rundown of what they did to come up with this paper.

iii. KNN

The KNN Algorithm considers the closeness of characteristics. After finishing the training set, we use the test set to determine how to treat a given data point depending on how strongly we emphasized it in the training set. Classification using KNN yields a membership prediction for a given class.The votes of an item's closest neighbors are more important in determining its most frequent group classification. One advantage of the item is that it may be used to generate regression output, which can predict long-term traits. For the value of k, this estimate is calculated as the median of the estimates of its two nearest neighbors. The Euclidean distance is a common distance metric used in KNN algorithms.

iv. Improved Decision Tree

Decision tree classifiers are greedy, therefore you can't utilize a feature you choose in the first step even if it could improve your classification in later stages. Another possible cause of underperformance on new data is overfitting the training data. Due of this limitation, ensemble models are used. Ensemble approaches average the results of several models. The output from an ensemble model is often better than that of the individual components.

By combining information entropy according to various weights with coordination degree in rough set theory, they presented an enhanced ID3 method. With ID3, choosing the best feature is done using the information gain formula (eq. 3), however the calculation is complicated because of the logarithm in the procedure. A more straightforward method would allow for quicker decision tree construction, which is the premise upon which their study rested. This was accomplished by reducing the logarithmic equation of information gain to four simple mathematical operations, namely addition, subtraction, multiplication, and division, which greatly accelerated the process of choosing a decision tree. Approximation formula of Maclaurin formula was used to alleviate this issue in information entropy calculation in ID3 algorithm. A root node is chosen based on the attribute with the largest information gain, and shortly after, they utilized the method they inferred to calculate the information gain of all the characteristics. By using three different datasets, they contrasted the conventional ID3 technique with their suggested alternative. While their system outperformed ID3 in terms of structure and runtime for the first two small datasets, it was unable to outperform ID3 in terms of accuracy. With regard to execution speed, accuracy, and tree structure, their enhanced method surpassed the conventional ID3 on the third, big dataset.

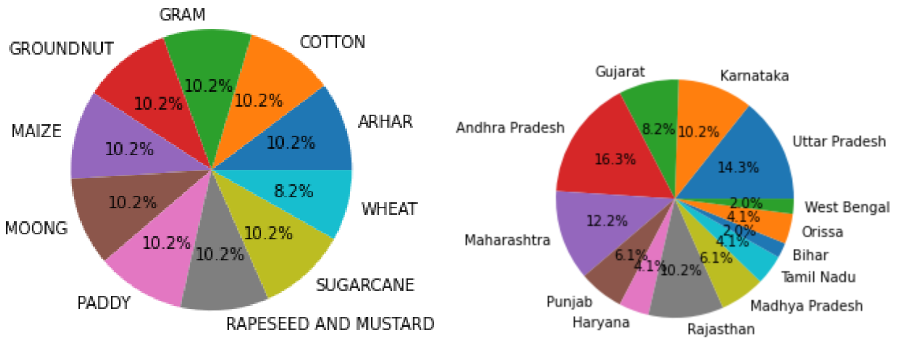


Figure 2. Grain growth percentage and pie chart of agriculture in different state with respect to peanut.

As a first phase in this study's data analysis, the data was classified according to several attributes and categories, including crop type, yield, condition, and so on. We run all the state-of-the-art approaches and the proposed methodology through our paces to ensure they provide accurate predictions, and the results are shown here.

TABLE. I.Percentage of Increase in Sugar Cane Grain

State	Yield (Quintal/Hectare)
Uttar Pradesh	448.89
Karnataka	986.21
Andhra Pradesh	757.92
Maharashtra	744.01
Tamil Nadu	1015.45

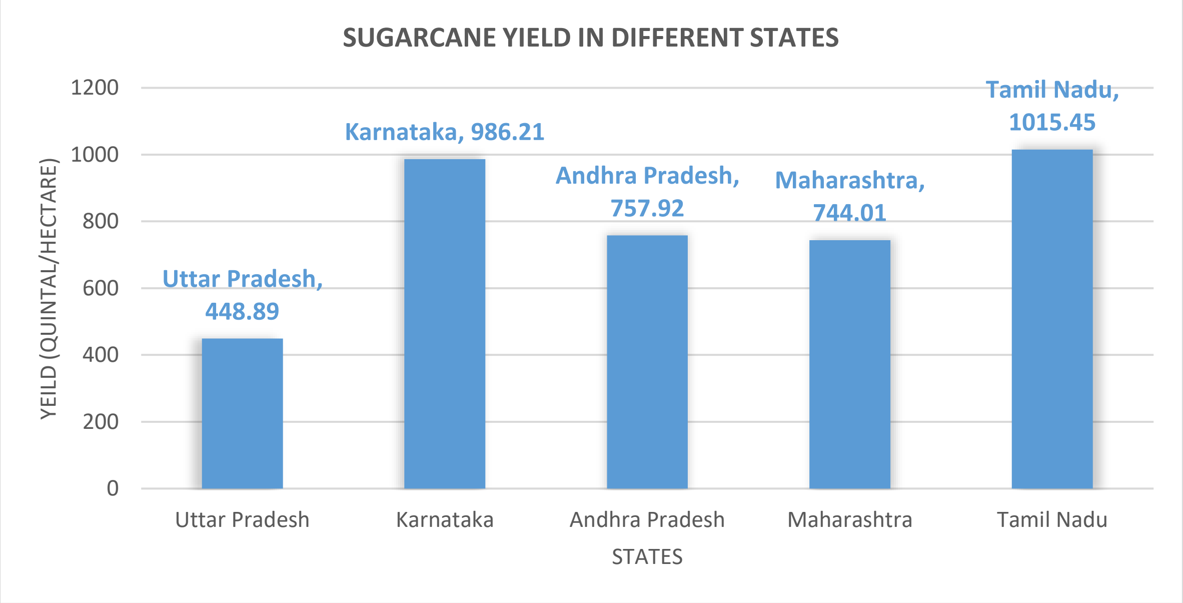


FIGURE 3. Sugar cane grain expansion rate

Table I and Figure 3 provide numerical and visual comparisons of the sugarcane yields, in quintals per hectare, from the five states. Uttar Pradesh produces a meager 448.89 quintal per hectare of sugarcane, in contrast to Tamil Nadu's high output of 1015.45 quintal per hectare.

Table. II.Andhra Pradesh's average agricultural yield per hectare (in quarts)

Crop	Yield (Quintal/ Hectare)
ARHAR	6.42
COTTON	17.83
GRAM	16.69
GROUNDNUT	11.97
MAIZE	42.68
MOONG	5.9
PADDY	56
SUGARCANE	757.92

Table II and Figure 4 provide numerical and visual comparisons of the different crops cultivated in Andhra Pradesh. The crop rotation includes arhar (toor), cotton, gram, groundnut, maize, moong, rice, and sugarcane. The yield is given in quintals per hectare. When considering the eight crops Sugarcane production is astronomically high when compared to moong dhal's output of 5.9 Q/H.

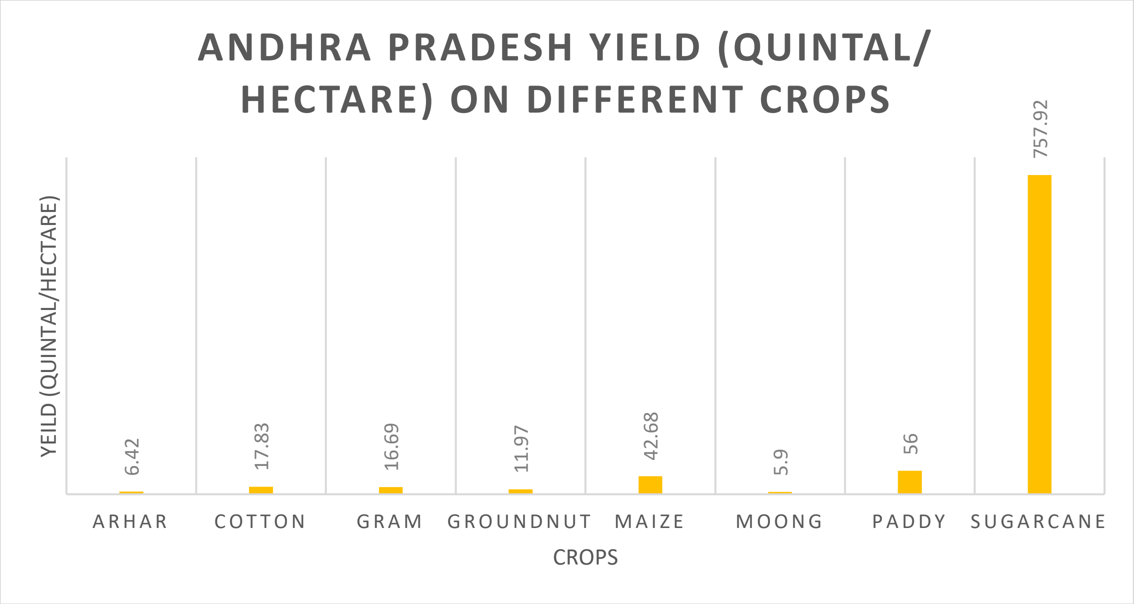


FIGURE4. Andhra Pradesh's average agricultural yield per hectare (in quarts)

Figure 4 shows a comparison of the total yield of all crops across all states, where sugarcane is the most prevalent crop and all others are produced in the lowest numbers. This yield is taken as the average total yield.

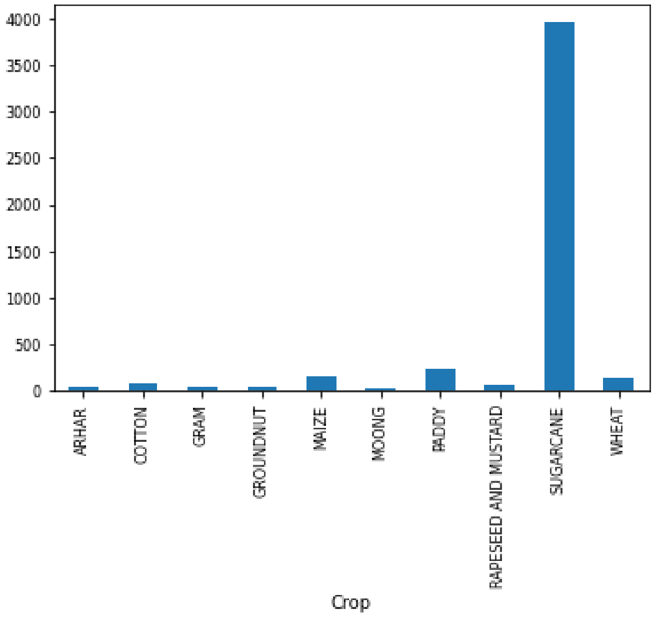


FIGURE 5. Yield in quintal per hectare

$$Accuracy = (TP + TN) / (TP + FP + TN + FN) \quad (8)$$

$$Specificity/Precision = TP / (TP + FP) \quad (9)$$

$$Sensitivity/Recall = TP / (TP + FN) \quad (10)$$

In this case, TP is when the expected occurrences are also positive. FP happens when things go wrong while people anticipate them to go well. When situations that should be bad really end up being negative, we say that they are TN. A false negative occurs when an anticipated negative outcome really turns out to be a positive one [7].

Table III. Matrix of Confusion Comparison of the proposed algorithm's parameters to those of similar approaches

Algorithm	TP	FN	FP	TN
SVM	78.22	21.78	18.62	81.38
RF	81.85	18.15	19.12	80.88
KNN	80.08	19.92	17.22	82.78
IDT	88.54	11.46	13.69	86.31

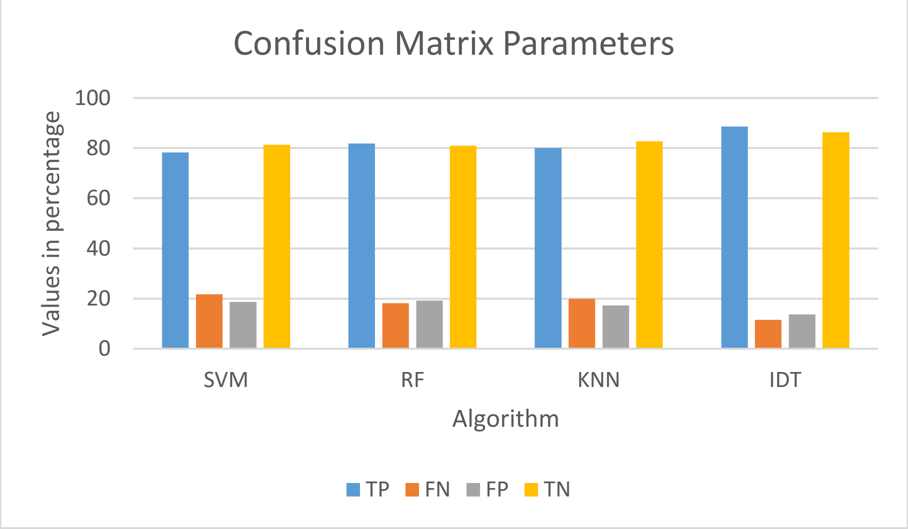


FIGURE 6. A Matrix of Confusion Comparison of the proposed algorithm's parameters to those of similar current approaches

TABLE IV. Comparison of the Proposed Algorithm's Validation Parameters to Currently Used Methods

Algorithm	Precision	Recall	Fscore	Specificity
SVM	80.77	78.22	79.48	81.38
RF	81.06	81.85	81.45	80.88
KNN	82.30	80.08	81.18	82.78
IDT	86.61	88.54	87.56	86.31

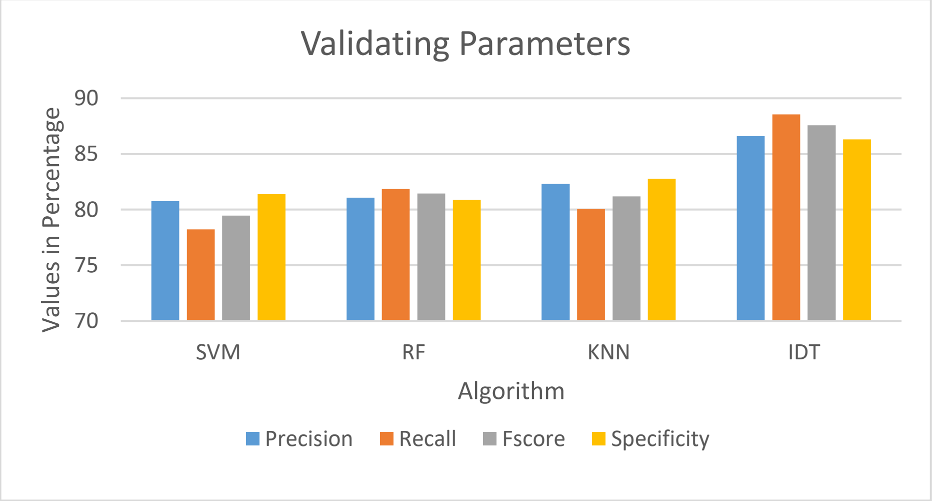


FIGURE 7. Examining the suggested algorithm's parameters against those of similar approaches

TABLE V. Validity evaluation of new and established approaches

Algorithm	Accuracy
SVM	79.80
RF	81.37
KNN	81.43
IDT	87.43

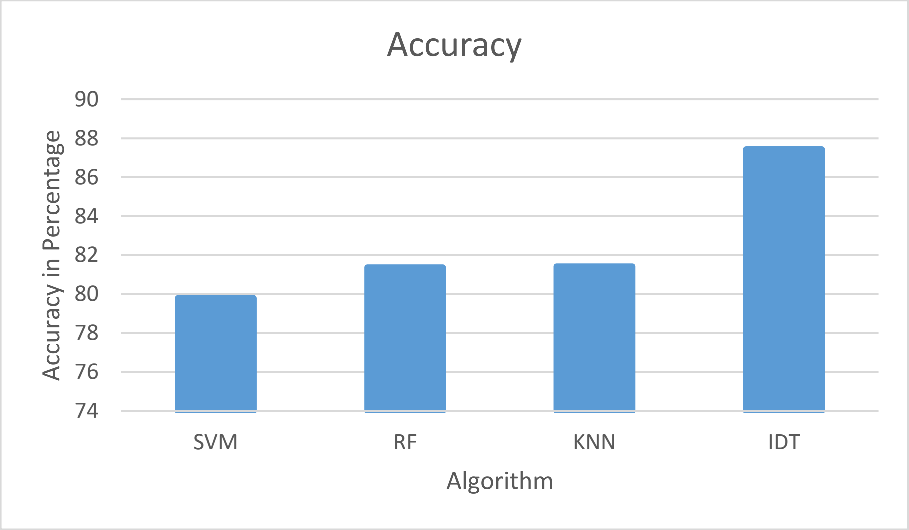


FIGURE 8.Evaluate the effectiveness of new and established approaches

Tables III, IV, and V, with the accompanying figures 6, 7, and 8, show a numerical and graphical comparison of the proposed algorithm's and the existing methods' accuracy. They also compare the proposed algorithm's Confusion Matrix parameters to the existing methods', as well as its Validation Parameters to the existing methods'. With respect to specificity, accuracy, recall, and F-score, the proposed method is clearly superior. The number of false positives and negatives is likewise the lowest for this. All other approaches pale in comparison to the proposed IDM method in terms of accuracy.

IV. CONCLUSION

This strategy is being proposed as a means to address the concerning increase in farmer suicides and provide them with assistance in bettering their financial circumstances. If farmers use our Crop Recommender system, they will have an easier time forecasting crop yields and choose the best crops to grow. On top of that, it tells you when it's best to sprinkle fertilizer. Appropriate machine learning tools were used for data collection, analysis, and training on relevant datasets. This article describes the suggested model in depth, detailing how it improves upon the Decision tree algorithm. A comparison is made between the suggested model and SVM, RF, and KNN.

REFERENCES

1. Sangeeta, S. G. (2020). Design and implementation of crop yield prediction model in agriculture. *International Journal of Scientific & Technology Research*, *8*(1), 544-549.
2. Sharma, B., Yadav, J. K. P. S., & Yadav, S. (2020, June). Predict crop production in India using machine learning technique: a survey. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 993-997). IEEE.
3. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., &Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.
4. Vapnik, V. Support vector machine. Mach. Learn, 20, 273–297.65. Suykens, J.A.K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. Neural Process. Lett. 1999, 9, 293–300, 1995.
5. Chang, C.; Lin, C. LIBSVM: A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol. (2013), 2, 1–39. [CrossRef]
6. Mshou, D.; Bravo, C.; West, J.; Wahlen, S.; McCartney, A.; Ramon, Automatic detection of “yellow rust” in wheat using reflectance measurements and neural networks. Comput. Electron. Agric., 44, 173–188, 2004. [CrossRef]
7. Moshou, D.; Bravo, C.; Oberti, R.; West, J.; Bodria, L.; McCartney, A.; Ramon, H. Plant disease detection based on data fusion of hyperspectral and multispectral fluorescence imaging using Kohonen maps. Real-Time Imaging, 11, 75–83, 2005.
8. Mitchell, T. (1997). Machine Learning. McGraw Hill. p. 2. ISBN 978-0-07-042807-2.Rakesh Kaundal1, Amar S Kapoor2, and Gajendra P, S Raghava*1. Machine learning techniques in disease forecasting: a case study on rice blast prediction". BMC Bioinformatics, 7:485, 2006.
9. Asadollahi, H., Kamarposhty, M.S., Teymoori, M. M. Classification and evaluation of tomato images using several classifiers. In: Paper presented at the Computer Science and Information TechnologySpring Conference, 2009. IACSITSC'09.International Association of, pp.471–474, 2009.
10. Moshou, D.; Bravo, C.; Wahlen, S.; West, J.; McCartney, A.; De Baerdemaeker, J.; Ramon, H. Simultaneous identification of plant stresses and diseases in arable crops using proximal optical sensing and self-organizing maps. Precis. Agric., 7, 149–164, 2006.