Journal of Advances in Science and Technology
Vol. 21, Issue No. 1, March-2024, ISSN 2230-9659

DOI: https://doi.org/10.29070/bs9vmv33

Check for updates

Propose and implement a Rule-Based System to Predict Crop Yield Production

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

1. Research Scholar, Awadhesh Pratap Singh University, Rewa, Madhya Pradesh, India
ashwanikhajuraho@gmail.com ,
2. Professor, Department of Computer Science, Awadhesh Pratap Singh University, Rewa, Madhya Pradesh, India

Abstract: Accurate meteorological forecasting plays a pivotal role in agricultural decision-making, particularly in determining crop yield potential and optimizing agricultural practices. This study investigates the application of data mining techniques in meteorological forecasting to enhance crop yield prediction accuracy Preliminary findings suggest that data mining techniques, including machine learning algorithms, neural networks, and ensemble methods, offer significant potential for improving the accuracy and reliability of meteorological forecasts for crop yield prediction. In conclusion, this study underscores the potential of data mining techniques in improving meteorological forecasting for crop yield potential assessment.

Keywords: Predict, Crop, Yield, Production and Climate

X

INTRODUCTION

There are many forces at work in agriculture, but one of the most important is climate change. Farmers are already feeling the effects of this global warming, and the weather is a major predictor of crop yields. Crop yields, animal populations, and ecological systems will all be more or less impacted by climate change, the precise nature of which varies around the globe. An increasing number of people are worried about the future of food production on a worldwide scale because of the predicted and actual consequences of climate change on farming and food safety. Increases in average global temperature, shifts in precipitation patterns, desertification, water body drying out, flooding, and deteriorating soil conditions all have an influence on agricultural production, crop yield, and the viability of agricultural products. Climate change also promotes the growth of pests and diseases.

Because agricultural production is a bio-socio-system including the interplay of soil, air, water, and crops, a complete model is required, which can be achieved only by classical engineering knowledge. The United Nations Food and Agriculture Organisation defines crop forecasting as the practice of estimating future crop output and yields many months before harvest. Meteorological, agro-meteorological, soil, remotely sensed, and agricultural statistics data are the cornerstones of crop forecasting philosophies. Several indices are developed from meteorological and agronomic data that are considered important factors in predicting crop production. These include crop water satisfaction, surplus and excess moisture, average soil moisture, and more. The linear regression model describes the mathematical connections that are inherent to the dataset derived from earlier trials. If the data used to build and test the model is comprehensive, this strategy may provide findings in a variety of contexts. However, there is a lack of full and sparse information in agricultural statistics. This constraint is why the linear regression method is the

de facto standard for yield prediction across expansive areas.

Multivariate regression models are the most studied crop-yield-weather models statistically. Data mining is a process that tries to discover interesting and useful patterns in data for farmers. The inability to accurately anticipate yields is a typical issue.

Both 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.

LITERATURE REVIEW

Kamir (2020) Using climatic records and NDVI time-series data analysis, uses machine learning to forecast wheat yield, capitalising on a large training dataset of high-resolution yield maps. Research shows that the most accurate method is support vector regression using radial basis functions. The results of this study demonstrated that non-linear models outperformed linear ones, whereas ensembles failed to outperform individual models.

Nevavuori (2019) The model for crop yield prediction is built using Convolutional Neural Networks (CNNs), a deep learning technique that has shown to be very effective in image classification applications. The model is trained using RGB and NDVI data collected from UAVs. We evaluate how changing several CNN parameters, such as network depth, training technique, hyperparameter tuning, and regularisation approach, affect prediction efficiency. During the growth season, CNN and L2 regularisation MAE and MAPE were computed using the Adadelta training method. According to the findings, CNN architecture outperformed NDVI data when fed RGB data.

Sharma, Negativeya (2019) In this work, we surveyed data mining approaches that have been used to estimate agricultural yields. For the purpose of predicting agricultural yields, it summarises the works of the algorithm that previous writers have used. Researchers find it highly useful for gathering data on the present state of data mining methods and applications applicable to the agricultural sector, as it consolidates the work of several writers in one location.

Chlingaryan (2018) This paper discusses research developments conducted in the past 15 years on machine learning techniques for crop yield prediction and nitrogen management. The outcome of this research shows that ML techniques and sensing technologies provide cost-effective and solutions for better decision making and crop estimation.

Ravneet Kaur Sidhu et al (2020) Many people throughout the globe rely on rice as a staple crop. Because of the high-water consumption required for its production, effective water management is essential for the

long-term viability of this resource. But there is a severe lack of information on the water use of rice irrigation. Predicting the daily watering plan of rice has been done using typical machine learning approaches in this work. In order to train and optimise the models, we utilise data from 2013 to 2015. The models are tested using data from 2016-2017. Feature selection using correlation criteria helps reduce the number of input parameters from 26 to 11 in the end. Based on the predicted weather conditions, the models calculated the amount of water needed by the crops. Compared to other models, Adaboost's average accuracy in forecasting the irrigation schedule was 71%, demonstrating consistent well-performance.

RESEARCH METHODOLOGY

The data used in this suggested study comes from 28 separate districts in Chhattisgarh and covers the period from 2000 to 2015. Total precipitation The Meteorological Department likewise gathered data for the same 28 districts over a period of 15 years. One kind of data mining is data preparation, which entails making raw data more understandable. The purpose of preprocessing is to reduce the dataset's dimensionality. The process of dimensionality reduction is described Let E be a dataset that contains n data vectors and is represented as a matrix with dimensions n×m. yi where (i ∈ {1, 2,...n}) as a function of dimension DI. After that, the data is sent into Matlab's curve fitting tool for principal component analysis (PCA) to determine which independent and dependent variables fit together best. As a regression tool, Support Vector Machine maintains all the key characteristics (maximum margin) that characterised the method. With a few key differences, the Support Vector or Regression (SVR) and the Support Vector Machine (SVM) share many of the same ideas for categorization.

RESULT

The Effect Of Rainfall On Crop Yield

Growing cotton uses around 3% of the world's irrigation water, although only 2.2% of the world's arable land is actually dedicated to this crop. So, a lot of irrigation water goes into growing cotton. Unless you're using a dry farming cotton production strategy, the water needed for crops comes from both natural precipitation during harvest and artificial irrigation in the months thereafter.

The Food and Agriculture Organisation (2012) defines water usage as the amount of water actually evaporated from the soil. Two distinct processes, evaporation and transpiration, work together to form evapotranspiration, which is the mechanism by which water is lost from the soil surface and consumed by the crop. It gives a quantification of the overall water consumption throughout crop growth in the field or the water loss from extraction to crop supply. ETa ranges from 410 to 780 mm seasonally, with the exact amount dependent on the irrigation technique and the amount of deficit irrigation (of precipitation). A sufficient amount of water is needed to enable 700 ml of evapotranspiration (the sum of soil evaporation and transpiration) in order for cotton to achieve satisfactory yields. Soil type and other soil properties; weather conditions (such as precipitation, relative humidity, wind speed, etc.); and other environmental elements are the determinants of ETa.

There are three main aspects of water management: the amount of water extracted for irrigation, how efficiently that water is used to grow cotton, and the quality of the water, both when it is applied to the

crop and when it (possibly) leaves the farm through deep percolation or surface runoff. Figure 1 shows that the water need of cotton crops changes as they develop through their various phases. What really affects cotton crop yield is not the overall amount of precipitation, but rather the amount that falls throughout the three distinct phases of growth: germination, fruiting, and maturity.

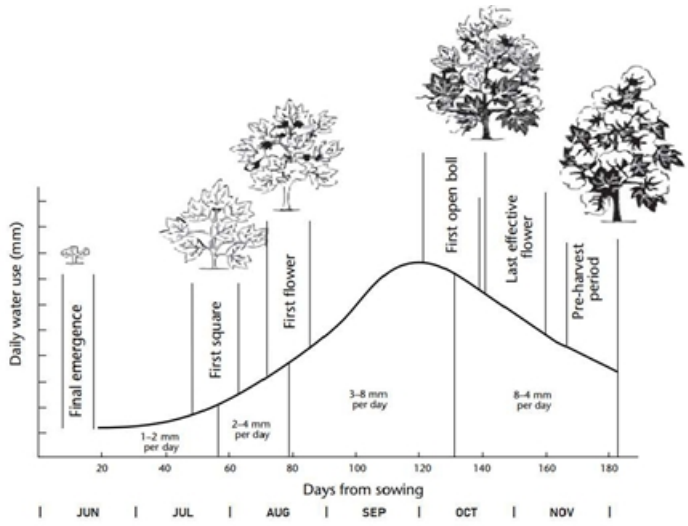

Figure 1: Graph of optimum water needs of cotton during growth period

Reduced photosynthetic activity and increased leaf senescence are symptoms of water stress, which is caused by a lack of water. A reduction in blooming is caused by the significant shedding of tiny squares caused by drought stress. Although boll abscission occurs in response to water stress in the first fourteen days after anthesis (the beginning of a plant's blooming season, beginning with the opening of the flower bud), neither flowers nor huge squares or balls shed very often. Thus, it is not uncommon for immature plants to keep flowering even when subjected to extreme stress. Ballast size and seed weight are both affected by water stress in the 20-30 days after anthesis.

Severe heat in the afternoon is a common cause of water stress. When exposed to temperatures that are too high or too low, cotton plants and the fibres they produce suffer.

The impact of a delayed or failed monsoon may be mitigated if farmers own irrigation infrastructure. Both the vegetative and fruiting stages of a cotton crop are stunted when water is unavailable in the root zone. Total output is affected by growth stage, moisture shortage level, and water stress period.

Data Classification

The current research examined the impact of rainfall on cotton output by treating it as an independent variable. Therefore, cases of cotton production were compared to rainfall. Using secondary sources, we were able to compile ten-year rainfall averages for all three districts' talukas. After looking at the rainfall data from 2006 to 2015, the talukas in the three districts that were chosen were categorised as talukas with 'High,' 'Medium,' or 'Low' rainfall. Table 1 and 6.2 provide the specifics of the basis for categorization.

Table 1 Classification of Talukas on the basis of average rainfall

Rain classification			
	Balrampur	Bastar	Durg
Low	up to 93	up to 82	up to 116
Medium	94 to 132	83 to 126	117 to 141
High	Above 133	Above 127	Above 142

The Non-Parametric Test

Parametric tests, contrary to popular belief, may actually work really effectively with non-normal continuous data provided that the sample size requirements are satisfied. Although nonparametric tests do not presume normally distributed data, they do entail additional, more challenging assumptions. A typical assumption for non-parametric tests comparing groups is that all groups' data must have the same spread (dispersion). Nonparametric tests may not provide reliable findings if the distributions of your groups are different. The term "distribution-free statistics" applies to non-parametric statistics that do not make any assumptions on the population's distribution. As a result, data with a large range of variation may be readily accommodated. These distribution-free tests are applicable to both quantitative and qualitative data, in contrast to parametric statistics.

The Wilcoxon Signed Rank Test

We used either the Wilcoxon signed rank test or the Wilcoxon matched pairs test to determine if rainfall had an influence on cotton yield. It is recommended to use the Wilcoxon signed rank test since the data is constituted of definite scores. When the same people are tested under two distinct circumstances, it works well as a repeated measure design. The yield and rainfall data sets were tested for the same taluka. The paired t-test in parametric form has a nonparametric counterpart. The degree to which the two pairings vary is taken into account by this test. Compared to the basic sign test, it makes more use of the data included in the score sets. It is thought to be more accurate than the sign test due to the fact that it utilises more information. For this test to work, it is essential that the two sets of scores being considered be connected and on an ordinal scale.

In the first set of numbers, we can see the ten-year average rainfall over all three districts and talukas. As a duo, they used cotton yield data from several years. We looked at the rainfall and yield statistics side by side to see how much of a difference there was. Table 2 displays the findings.

Table 2 Wilcoxon Signed-Rank Test Result for rainfall

Wilcoxon Signed-Rank Test					
Treatment 1	Treatment 2	*Sign*	*Abs*	*R*	*Sign R*
150.12	680	-1	529.88	23	-23
145.85	879	-1	733.15	30	-30
96.67	598	-1	501.33	19	-19
65.27	691	-1	625.73	29	-29
201.63	561	-1	359.37	8	-8
152.55	597	-1	444.45	14	-14
104.7	643	-1	538.3	25	-25
180.63	599	-1	418.37	11	-11
103.82	234	-1	130.18	1	-1
155.2	717	-1	561.8	26	-26
152.52	467	-1	314.48	6	-6
141.83	440	-1	298.17	4	-4
72.17	470	-1	397.83	10	-10
95.22	594	-1	498.78	18	-18
155.13	639	-1	483.87	17	-17
146.68	667	-1	520.32	21	-21
153.3	758	-1	604.7	28	-28
180.87	702	-1	521.13	22	-22
74.9	613	-1	538.1	24	-24
145.2	655	-1	509.8	20	-20
188.63	766	-1	577.37	27	-27
155.07	389	-1	233.93	3	-3
115.33	508	-1	392.67	9	-9
88.6	558	-1	469.4	16	-16
137.42	450	-1	312.58	5	-5
106.77	550	-1	443.23	13	-13
98.47	531	-1	432.53	12	-12
170.87	487	-1	316.13	7	-7
130.93	351	-1	220.07	2	-2
125.43	590	-1	464.57	15	-15

There was a significant difference between the rainfall and cotton yield pairings according to a Wilcoxon signed rank test; with n = 30, Z = -4.7821, and $p < 0.05$. There is a noticeable disparity in size between the two groups.

Pearson's Correlation Coefficient

Also, we tried to figure out how strongly the two variables were related to each other. The Pearson's Correlation Coefficient was used for this objective. To find out how closely related or associated two continuous variables are statistically, statisticians use Pearson's correlation coefficient. Because it is based on the concept of covariance, it is renowned as the finest way to measure the relationship between variables of interest. It reveals not just the direction of the link but also its strength, sometimes called a correlation.

The likelihood that you would have obtained the present result under the null hypothesis, when the correlation coefficient is 0, is known as the P-value. The correlation coefficient is deemed statistically significant if the probability is less than the standard 5% ($P < 0.05$). Exactly what magnitude of connection is deemed high, moderate, or weak is not defined by any rule. The values can never be more than -1 (very negative correlation) or +1 (very positive correlation). Values around 0 indicate a weak or nonexistent linear connection. We normally consider correlations over 0.4 to be quite high for this kind of data. Moderate correlations are those between 0.2 and 0.4, while weak correlations are those below 0.2. Table 3 displays the findings.

Table 3 Pearson Correlation Coefficient Result for rainfall

Pearson Correlation Coefficient						
x value	y value	$X - M_x$	$Y - M_y$	$(X - M_x)^2$	$(Y - M_y)^2$	$(X - M_x)(Y - M_y)$
150.12	680	17.061	100.533	291.066	10106.951	1715.166
145.85	879	12.791	299.533	163.601	89720.218	3831.231
96.67	598	-36.389	18.533	1324.184	343.484	-674.416
65.27	691	-67.789	111.533	4595.394	12439.684	-7560.77
201.63	561	68.571	-18.467	4701.936	341.018	-1266.272
152.55	597	19.491	17.533	379.886	307.418	341.736
104.7	643	-28.359	63.533	804.252	4036.484	-1801.763
180.63	599	47.571	19.533	2262.968	381.551	929.214
103.82	234	-29.239	-345.467	854.939	119347.218	10101.215
155.2	717	22.141	137.533	490.209	18915.418	3045.08
152.52	467	19.461	-112.467	378.718	12648.751	-2188.676
141.83	440	8.771	-139.467	76.925	19450.951	-1223.216
72.17	470	-60.889	-109.467	3707.511	11982.951	6665.352
95.22	594	-37.839	14.533	1431.815	211.218	-549.932
155.13	639	22.071	59.533	487.114	3544.218	1313.94
146.68	667	13.621	87.533	185.523	7662.084	1192.262
153.3	758	20.241	178.533	409.685	31874.151	3613.634
180.87	702	47.811	122.533	2285.86	15014.418	5858.4
74.9	613	-58.159	33.533	3382.508	1124.484	-1950.276
145.2	655	12.141	75.533	147.396	5705.284	917.025
188.63	766	55.571	186.533	3088.099	34794.684	10365.782
155.07	389	22.011	-190.467	484.469	36277.551	-4192.298
115.33	508	-17.729	-71.467	314.329	5107.484	1267.056
88.6	558	-44.459	-21.467	1976.632	460.818	954.394
137.42	450	4.361	-129.467	19.015	16761.618	-564.561
106.77	550	-26.289	-29.467	691.129	868.284	774.659
98.47	531	-34.589	-48.467	1196.422	2349.018	1676.43
170.87	487	37.811	-92.467	1429.647	8550.084	-3496.226
130.93	351	-2.129	-228.467	4.534	52197.018	486.482
125.43	590	-7.629	10.533	58.207	110.951	-80.362
		Mx: 133.059	My: 579.467	Sum: 37623.972	Sum: 522635.467	Sum: 29500.289

The variables were positively associated with each other, with an R-value of 0.2104 and a p-value less than.05.

Figure 2 shows the dispersion graph of 30 data pairs, and Figure 3 shows the average rainfall over a ten-year period. The PCC for these pairings was found to be 0.2104 ($p<0.05$). A straight line is shown by the scattered layout, despite the dots being all over the place.

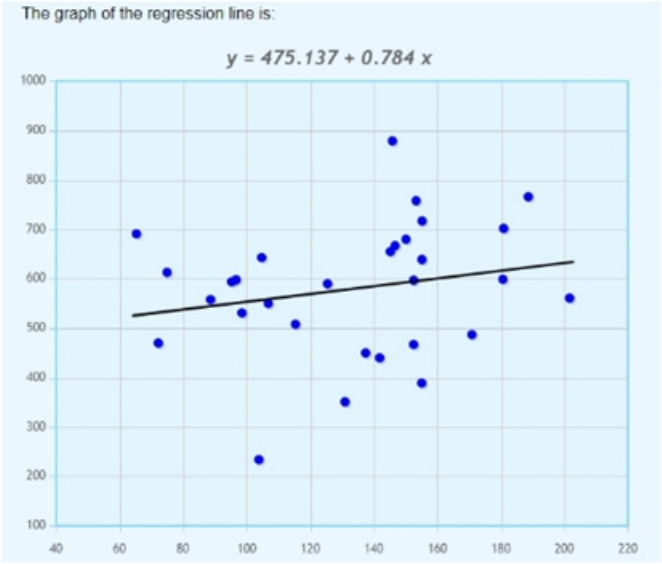

Figure 2 Scattered graph for rainfall and yield correlation

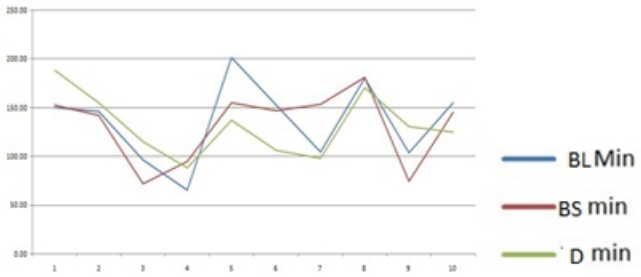

Figure 3 Average rainfalls for ten years from 2006 to 2015

Weka Classification and Calculations

Data mining (machine learning) classification is a method for predicting which groups data examples will belong to. Finding a model for a class attribute based on the values of other characteristics and then accurately assigning test data to those classes is the task at hand. Model creation, or specifying a collection of established classes, is the first stage in classification. The second step is to use that model for prediction, or categorizing future or unknown things.

There are two types of classifiers in WEKA: supervised and unsupervised. None of the classifiers such as sluggish, tree, rules, and naïve fall outside of these specific groups. To further improve the precision of classifiers trained with different assemblers, meta classifiers are also available. In Figure 4, we can see the

stages of the WEKA classification algorithm.

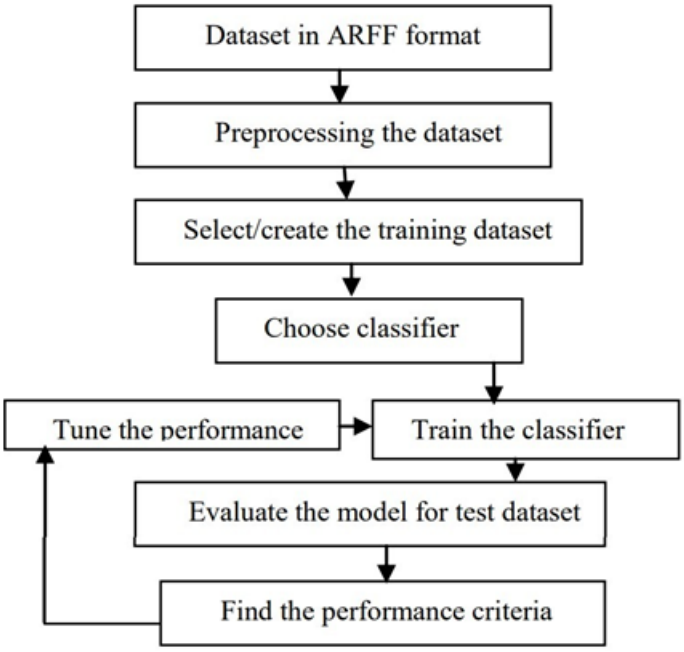

Figure 4 WEKA Classification Steps (Ref. Shweta Srivastava (2014))

Using WEKA's cross validation features and the goal variable "Yield," several classification algorithms were applied to rainfall data. Multiple methods were used for the assessment, including Multilayer Preceptron (MLP), Additive Regression (AR), Kstar, Sequential Minimal Optimisation (SMO), and Gaussian Processes (GP).

In GP, the distribution is over the mean and covariance functions, and the classifier function does not need hyperparameter adjustment; it is a probabilistic approach to regression and classification. ML-P utilises back propagation, a supervised learning method, to classify cases. It is a feed-forward multi-layer neural network classifier. Many applications rely on MLPs for tasks such as pattern recognition, estimation, prediction, and classification. The Support Vector Machine (SVM) is used for regression in the SMOreg method. A line that effectively divides the training data into classes is identified by the support vector machine. Meta classifier additive regression (AR) improves a regression base classifier's performance. It's not just an algorithm; it's based on a Bayesian probability model. For the rainfall and yield data, Table 4 displays the computed coefficients and Root Mean Square Errors (RMSE).

Table 4 WEKA algorithms result for rainfall

WEKA Algorithm	Correlation co-efficient	RMSE	RMSE Cross Valid
Gaussian Processes	0.2743	1.169307	1.318599
MLP	0.2881	1.735576	1.448596
SMOreg	0.2743	1.070578	1.366652
Kster	0.2752	1.52543	1.544968
Additive Regression	0.2507	1.778639	1.762545

The table shows that the RMS errors, correlation coefficients, and features of the algorithms that were tested for rainfall and cotton crop output varied. We focused on the correlation coefficient and root-mean-squared errors (RMSEs). In the cross validation run, GP had the best performance across all three metrics, with a correlation of 0.2743, an RMSE of 1.169307, and an RMSE of 1.318599.

CONCLUSIONS

Using a dimension reduction method to narrow the focus of the collected data, this study offered a machine learning model for predicting crop yields; this model would exclude information that may compromise the quality of the model. We employed principal component analysis (PCA) to preprocess the input dataset, and then we used the K-medoid clustering approach to get better predictions. Estimates obtained with WEKA make use of the linear regression method. When day-based forecasting is sought for, satisfactory outcomes are not achieved. But when we wanted the monthly average, we got the same results. So, it is clear that using a monthly average in the linear regression method for weather forecasting in WEKA produces good results.

References

1. Kamir, E. W. (2020). Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. Kamir, E., Waldner, F., & Hochman, Z. Estimating ISPRS Journal of Photogrammetry and Remote Sensing, 160, 124–135
2. Nevavuori, P. N. (2019). Crop yield prediction with deep convolutional neural networks. Computers and Electronics in Agriculture, 1-9.
3. Chlingaryan, A. S. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, 61–69.
4. Sidhu, Ravneet & Kumar, Ravinder & Rana, Prashant. (2020). Machine learning based crop water demand forecasting using minimum climatological data. Multimedia Tools and Applications. 79. 13109-13124. 10.1007/s11042-019-08533-w.
5. D, Dr & Kamate, Nikhil & Gawade, Om & Matruprasad, Pinaki & Sai, Vamsi. (2024). Soil Analyser - Revolutionizing Agriculture through Wireless Sensor Networks and Machine Learning. International Journal for Research in Applied Science and Engineering Technology. 12. 85-88. 10.22214/ijraset.2024.58248.

6. Afrin, Sadia & Khan, Abu & Mahia, Mahrin & Ahsan, Rahbar & Mishal, Mahbubur & Ahmed, Wasit & Rahman, Mohammad. (2018). Analysis of Soil Properties and Climatic Data to Predict Crop Yields and Cluster Different Agricultural Regions of Bangladesh. 80-85. 10.1109/ICIS.2018.8466397.

7. Sakthipriya, S. & Naresh, R.. (2023). Precision agriculture: crop yields classification techniques in thermo humidity sensors. Optical and Quantum Electronics. 56. 10.1007/s11082-023-05907-1.

8. Gowda, Shruthi & Reddy, Sangeetha. (2020). Design And Implementation Of Crop Yield Prediction Model In Agriculture. International Journal of Scientific & Technology Research. VOLUME 8,. 544.

9. Maqsood, Junaid & Farooque, Aitazaz & Abbas, Farhat & Esau, Travis & Wang, Xiuquan (Xander) & Acharya, Bishnu & Afzaal, Hassan. (2022). Application of Artificial Neural Networks to Project Reference Evapotranspiration Under Climate Change Scenarios. Water Resources Management. 36. 1-17. 10.1007/s11269-021-02997-y.

10. Afzaal, Hassan & Farooque, Aitazaz & Abbas, Farhat & Acharya, Bishnu & Esau, Travis. (2020). Computation of Evapotranspiration with Artificial Intelligence for Precision Water Resource Management. Applied Sciences. 10. 10.3390/app10051621.

11. Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal, Crop yield prediction using machine learning: A systematic literature review, Computers and Electronics in Agriculture, Volume 177, 2020, 105709, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2020.105709.

12. Mohanadevi M et al (2018) " A Study on Various Data Mining Techniques for Agriculture Crop Yield Prediction", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.5, Issue 9, page no.797-802, September-2018, Available: http://www.jetir.org/papers/JETIR1809626.pdf

13. C, Mithra & Suhasini, A. (2023). Fertilizer type and quantity recommendation to increase oilseed crops yield prediction with inorganic fertilizers using machine learning algorithms. Ecology, Environment and Conservation. 29. 358-372. 10.53550/EEC. 2023.v29i02s.059.

14. Shah, Sayed. (2021). Machine Learning based Crop Recommendation System for Local Farmers of Pakistan. Revista Gestão Inovação e Tecnologias. 5735-5746. 10.47059/revistageintec. v11i4.2613.

15. Katuru, K. & Surapaneni, Ravi & Dasari, Suresh. (2020). Predicting Crop yield and Effective use of Fertilizers using Machine Learning Techniques. International Journal of Innovative Technology and Exploring Engineering. 9. 1288-1292. 10.35940/ijitee. G5911.059720.