Autonomous Crop Prediction System using Machine Learning

Praveen Kumar Maduri
Amity University

Dhiman, Preeti
Galgotias college of engineering and technology (AKTU), India

Srivastava, Rishabh
Galgotias college of engineering and technology (AKTU), India

Singh, Riya
Galgotias college of engineering and technology (AKTU), India

https://doi.org/10.5109/7183439

出版情報：Evergreen. 11 (2), pp.1299-1304, 2024-06. 九州大学グリーンテクノロジー研究教育センター
バージョン：

Autonomous Crop Prediction System using Machine Learning

Praveen Kumar Maduri[1], Preeti Dhiman[2], Rishabh Srivastava[2*], Riya Singh[2]
[1]Amity University, Kolkata
[2] Galgotias college of engineering and technology (AKTU), India

Email: 16rishabh6@gmail.com

(Received September 14, 2022; Revised April 8, 2024; Accepted June 14, 2024)

Abstract: Agriculture shows its considerable contribution to the GDP of India. The common citizens of India turn on directly or indirectly on farming for their living. Prediction of crop in agricultural field is based on different input variables. Here we proposed a model in which different input parameters are analyzed using different machine learning algorithms. Automation based crop production system in agriculture field help the farmers to enhance the yield of crop by introducing a best crop for the agriculture field. This model helps in analyzing the best suitable machine learning algorithm by monitoring the accuracy with different training and testing ratios.

Keywords: Machine Learning, Prediction analytics, Decision Tree, kNN, Random Forest, Naive Bayes Gradient Boosting

1. Introduction

In our country, the automatic concept of farming the field is not completely evolved and farmers still uses the old traditional techniques for growing plants which is not effective to lead to a great yield and at the same time had a high risk of lack of success. Economists have proved that agriculture and farming are the precursors of economic development which contribute immensely to its development e.g., by supplying wage goods to industrial workers, by transferring surplus from agriculture to finance, for industrialization, by using the product of industry as investment for the agricultural sector and by transferring surplus labor from agriculture to industrial jobs, contributing to the development of the country. Our country is facing many difficulties in the field of agriculture, many new technologies has been developed still India is facing many challenges like lack of water during early crop development can result in reduced production or failure of the entire production. Therefore, we may leverage cutting-edge technologies like machine learning and artificial intelligence to lessen these obstacles. With the aid of artificial intelligence and data prediction, agricultural output can be increased. AI robots are also being developed to undertake other agricultural jobs including harvesting and monitoring soil health. Additionally, machine learning algorithms are capable of analysing and forecasting environmental effects like climate change. Basically, we can predict the crop which is basically suitable for the field and environment, with the help of sensors and Machine Learning. The machine learning algorithms have not been widely used for crop classification, and as to our knowledge, their performance in this type of application has not been thoroughly compared[1-3]. With the aid of the Global Positioning System, sensors are utilised to measure the necessary characteristics of the soil (GPS). Crop sensors are also utilised to instantly control equipment for variable rate application.

Farmers that adopt an advanced agricultural system mindset can find the right balance between productivity and post-harvest technologies. This will lead to help the farmers and also to the country by producing a large amount of crop or better growth.

Crop yield relies strongly on how effectively the basic land requirements can be utilized; land here refers to topography, soil type, soil nutrients, water content, sunlight, and all such factors related to crop growth[4-6]. The data collected from the agricultural field and the environmental surrounding are measured by sensors for the input parameter like NPK content, soil moisture, humidity, rainfall and pH of the soil etc. and these parameters are analyzed using different ml algorithms such as kNN, decision tree, random forest, Gaussian Naive Bayes and gradient boosting through different ratio of training and testing dataset. We are using different ratio of training and testing the model which us 70-30, 75-25 and 80-20. By the help of this, we can get the best accurate and suitable model to predict the crop for our agriculture fields.

2. Literature Review

J. Gholap et. al. suggested the Agricultural research can be beneficial by the use of technical advancement like automation and data mining [7-9]. It is young research in the field of agriculture. It aims that the classification of the soil type with the help of various Machine Learning Algorithms and the analysis of soil dataset usinh data mining technique. Another objective is to predict the attributes which remains untested with the help of regression techniques.

In research carried out by Zeel Doshi et. al. stated that the Indian population is dependent upon the agriculture directly or indirectly [8-11)]. The Indian farmers following the ancestral pattern without analysing the climatic conditions and soil conditions. So, there research represent an intelligent system known as AgroConsultant, which gives a prediction on the basis of climatic conditions, soil conditions, geographical location, and the sowing season.

Avinash Kumar el. al. described that the farmers are unaware of the minerals and the other parameters which are required by the crop, this can affect the crop as well as the farmers in both ways financially and mentally. Also, they stated that the pest and diseases can affect the growth of the crops [12-14)]. So, as a solution they are using some pest control techniques and also predicting the best suitable crop with the help of different machines learning algorithm i.e., SVM classification algorithm, Logistic Regression and Decision Tree algorithm.

In the research carried out by Nidhi H. Kulkarni and others, developed a system that uses ensembling technique of machine learning [15-17)]. Ensembling technique is made which integrate the number of ml models to anticipate the best crop according to the type of agriculture field. The different classifier used are first one is Random Forest, secondly the Naive Bayes, and the last one is Linear SVM. Each model gives their own data with its predicted accuracy. The labels of the class of every classifier are merged using the technique of majority voting. The ensembling technique is then used to classify the dataset of the soil into Kharif and Rabi.

The research which is carried out by the S. Pudumalar el. al. discussed about analyzing the biotic and abiotic factors in the agriculture using the data mining [18-21)]. The farmers didn't select the crop according to the soil conditions due to which they face a serious issue in the productivity. So, this kind of problem can be solved using the majority voting techniques using CHAID, Random Forest Classifier, Naive Bayes and K-nearest neighbors.

Kevin Tom Thomas el. al. designed a precision agriculture in which the soil property such as texture of soil, pH temperature, rainfall etc. of soil is used for observing that which crop is suitable for farming [22-25)]. This minimizes the incorrect prediction of crop and increases the productivity. Also, they are predicting the accuracy of different machine learning models. The various Machine Learning Algorithms used by them are K-nearest neighbors, Decision Tree, k-nearest neighbors with cross validation, Naïve Bayes and SVM.

3. Working

In this model, we have taken a open source dataset and found that it has 2200 rows and 8 columns present in it. We observe that the dataset has 22 unique classes and has 7 parameters, which is considered as the input and the predicted crop is the output. Then the input parameter is analyzed (data visualization) with the help of histogram plot. After data visualization, the correlation between the input parameters is checked in which we have found a strong relation between P(Phosphorous) and K(Potassium). Then we design three cases for model performance. In first case, the 70% data has been for the training the model and 30% data has been used for testing. According to that data split the accuracy of five algorithm i.e., i) Decision Tree (ii) Random Forest (iii) kNN (iv) Gaussian Naive Bayes and (v) Gradient Boosting, is checked. In second case, the 75% data is for the training and 25% data is for testing the model. According to that data split the accuracy of all the five algorithm is checked. In third case, the 80% data is measured for the training and 20% data has been used for testing. Similarly, data split the accuracy of five algorithm is measured.

When all the accuracies are measured then the best accuracy at different cases is observed, then the best combination (best accuracy of any model from the three training and testing cases) is taken for the prediction.

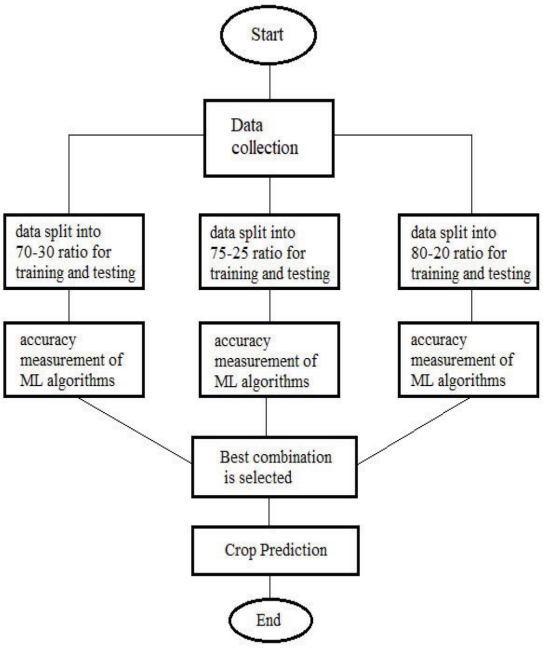

Start
Data collection
data split into 70-30 ratio for training and testing
data split into 75-25 ratio for training and testing
data split into 80-20 ratio for training and testing
accuracy measurement of ML algorithms
accuracy measurement of ML algorithms
accuracy measurement of ML algorithms
Best combination is selected
Crop Prediction
End

Fig.1: Flowchart

4. Methodology

4.1 Data Collection

The dataset consists various parameters of the soil and environment. There are various datasets available as an online open source that has been used here. So, in this model we are considering 22 types of crops i.e., rice, maize, chickpea, kidney Beas, pigeon peas, moth beans, mung beans, black gram, lentil, pomegranate, banana, mango, grapes, watermelon, musk melon, apple, orange, papaya, coconut, cotton, jute, coffee.

Figure 1 shows the number of instances of all the 22 crops. In this dataset, there are 2200 data values and all the crops has equal instances i.e., 100 instances for each crop.

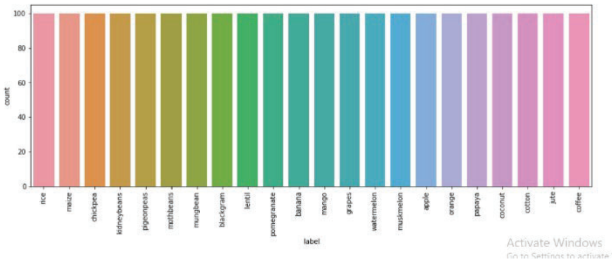
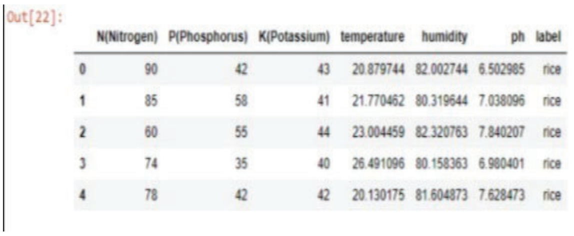


Fig. 2: Instances of crops

This dataset also consists some parameters which is considered for the crop recommendation. That parameters are NPK(Nitrogen-Phosphorus-Potassium) content of the soil, temperature and humidity, pH value of the soil and the amount of rainfall. Here, we can observe the dataset from Fig.3.



Out[22]:

	N(Nitrogen)	P(Phosphorus)	K(Potassium)	temperature	humidity	ph	label
0	90	42	43	20.879744	82.002744	6.502985	rice
1	85	58	41	21.770462	80.319644	7.038096	rice
2	60	55	44	23.004459	82.320763	7.840207	rice
3	74	35	40	26.491096	80.158363	6.980401	rice
4	78	42	42	20.130175	81.604873	7.628473	rice

Fig. 3: Dataset

Basically, soil is the main attribute for any crop. There are some techniques by which we can grow plants without soil but still soil is considered as the best suitable attribute for growing crops or any plant. So, we are observing NPK content as they are the macronutrients present in the soil which is very important for any crop. It fulfils all the necessary requirements of crop also it delivers better quality. So, the parameters we have measured are:

i)NPK content- NPK stands for Nitrogen-Phosphorus-Potassium. These are the basic essential macronutrients required for any crop.

ii)Temperature and humidity- Temperature and humidity is another an important aspect for the growth of plants or crop. In warm temperature, with low humidity plant started to go up into the transpiration state which reduces the growth of plant.

iii)pH value- Targeting the pH value of the soil is an important factor it ensures the bacteria present in the soil and also the availability of nutrients.

iv)Rainfall- The proper amount of rainfall, results in the better growth of the crop. The amount of rainfall should be balance. Excessive amount of rainfall can also damage the crops.

These parameters are the main aspect for any crop. So, the prediction is based on these parameters.

4.2 Preprocessor used for Analysing Dataset

The pre-processor steps which we have taken are as follows :

a) Library importing
b) Reading the dataset
c) Checking for unique values
d) Missing values checking
e) Plotting histograms of input parameters
f) Correlation between input parameters
g) Splitting of dataset

First, we have imported all the necessary libraries which are required to process all the steps and to run all the machine learning algorithm and then the dataset is being read.

Then the unique values are checked to analyse the number of classes present in the dataset. After checking unique values, the missing values is checked. Since, there is no missing values found, no further handling techniques is applied for the missing values. The histogram plotting is then done for the visualizing the dataset.

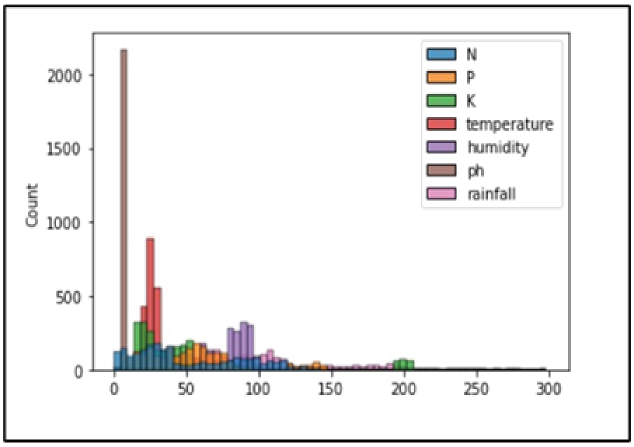


Fig. 4: Histogram plot of dataset

After data visualization the correlation between the variables (parameters) is done. We got the high correlation between the P (Phosphorus) and K(Potassium). Simply, we can observe it from Fig. 5. i.e., the correlation plot.

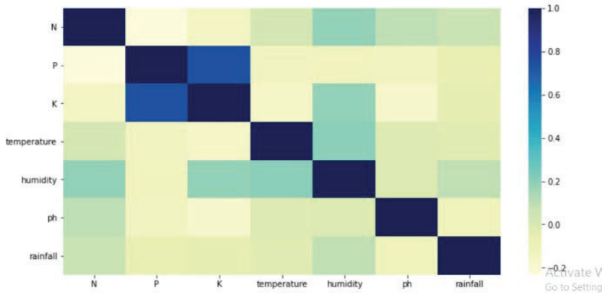


Fig. 5: Correlation Plot

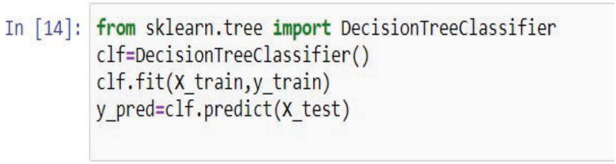
We have taken 3 cases for data splitting i.e. (i) 70% data for training and 30% data for testing
(ii) 75% data for training and 25% data for testing
(iii) 80% data for training and 20% data for testing.

According to these data split the appropriate machine learning model has been taken for the prediction of crop.

4.3 Accuracy of Models

i) Decision Tree Algorithm- This algorithm consists of parent node and child node. This algorithm works on the principle of flowchart. We have taken this algorithm because our data has multi classes. So, we are predicting the crop on the basis of parameters which is taken from the dataset. NPK content of the soil, Temperature and humidity, pH of the soil, amount of rainfall is considered as the input parameter for the algorithm and that data is splitted for the testing and training.

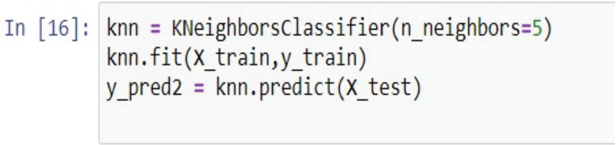
Now, the data is fitted into the Decision Tree Classifier for the prediction of crop.



```
In [14]: from sklearn.tree import DecisionTreeClassifier
         clf=DecisionTreeClassifier()
         clf.fit(X_train,y_train)
         y_pred=clf.predict(X_test)
```

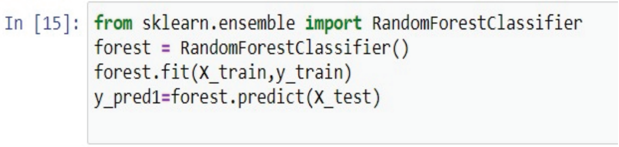
ii) kNN Algorithm- The kNN Algorithm is basically used for the classification. It is the simple supervised algorithm. kNN Algorithm is suitable for the multi class problem as our dataset has 22 classes. So, this algorithm is suitable for our system. The value of 'k' is calculated by the Euclidean formula. So, the value of 'k' is taken 1. Then data is splitted into the training and testing and fitted into the kNN Classifier.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



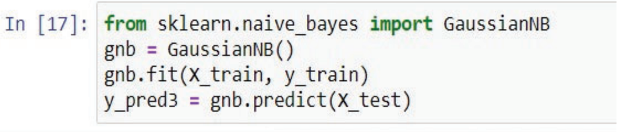
```
In [16]: knn = KNeighborsClassifier(n_neighbors=5)
         knn.fit(X_train,y_train)
         y_pred2 = knn.predict(X_test)
```

iii) Random Forest Algorithm- It is used for the problems having multiple classes and for the classification. This algorithm takes some samples from the dataset and a decision is made for every sample. Then a large number of decision trees is collected then the voting is performed to predict the result. So, our dataset is splitted into the training and testing and fitted into the Random Forest Classifier.



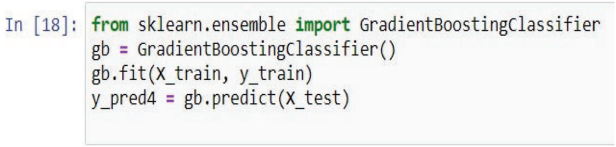
```
In [15]: from sklearn.ensemble import RandomForestClassifier
         forest = RandomForestClassifier()
         forest.fit(X_train,y_train)
         y_pred1=forest.predict(X_test)
```

iv) Gaussian Naive Bayes Algorithm- This Algorithm is used for the problems having multiple classes. This theorem is the extension of Naive Bayes Algorithm. It follows the Gaussian normal distribution and also it supports the continuous distribution of the data. As the dataset is continuously distributed and has multi class so we have used Gaussian Naive Bayes Algorithm. Then the dataset is splitted into the training and testing and fitted into the Gaussian Naive Bayes Classifier.



```
In [17]: from sklearn.naive_bayes import GaussianNB
         gnb = GaussianNB()
         gnb.fit(X_train, y_train)
         y_pred3 = gnb.predict(X_test)
```

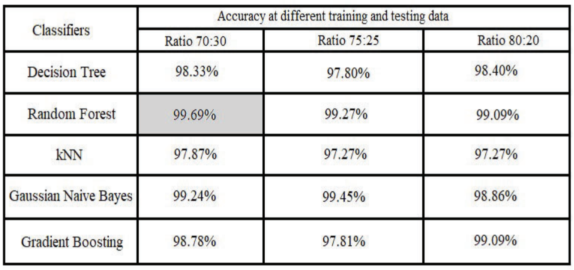
v) Gradient Boosting Algorithm- This algorithm usually used for the regression and classification. This theory predicts the output on the basis of arranging all the weak predictions. So, the dataset is splitted into the training and testing and fitted into the Gradient Boosting Classifier.



```
In [18]: from sklearn.ensemble import GradientBoostingClassifier
         gb = GradientBoostingClassifier()
         gb.fit(X_train, y_train)
         y_pred4 = gb.predict(X_test)
```

5. Outcome

Here, we have seen that by introducing different machine learning classifiers, we can clearly observe the difference between them. And by using different ratio of training and testing, we get to know the comparison between them that at which ratio, which classifier provides the best accurate result. So, the accuracy of all the five algorithms at different training and testing ratios is measured successfully.



Classifiers	Accuracy at different training and testing data		
	Ratio 70:30	Ratio 75:25	Ratio 80:20
Decision Tree	98.33%	97.80%	98.40%
Random Forest	99.69%	99.27%	99.09%
kNN	97.87%	97.27%	97.27%
Gaussian Naive Bayes	99.24%	99.45%	98.86%
Gradient Boosting	98.78%	97.81%	99.09%

Fig. 6: Tabular presentation of Accuracy

The above table represent the accuracies at different test cases of different classifiers.

So, we can observe at 70% training data and 30% testing data the accuracy of Decision trees, Random Forest Classifier, kNN Classifier, Gaussian Naive Bayes Classifier and Gradient Boosting Classifier is 98.33%, 99.69%, 97.87%, 99.24% and 98.78% respectively.

Similarly, at second case i.e. At 75% training data and 25% testing data the accuracy of Decision trees, Random Forest Classifier, kNN Classifier, Gaussian Naive Bayes Classifier and Gradient Boosting Classifier is 97.80%, 99.27%, 97.27%, 99.45% and 97.81% respectively.

Also, at 80% training data and 20% testing data the accuracy of Decision trees, Random Forest Classifier, kNN Classifier, Gaussian Naive Bayes Classifier and Gradient Boosting Classifier is 98.40%, 99.09%, 97.27%, 98.86% and 99.09% respectively.

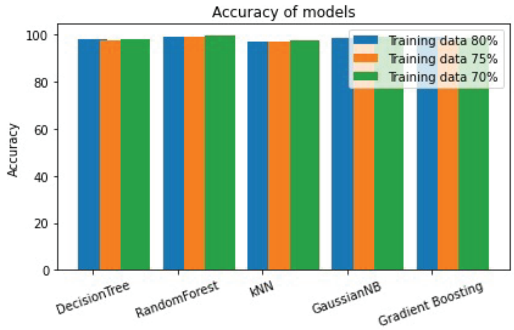


Fig. 7: Accuracy of models

So, we can see form the above data, we got the best accuracy at case first i.e., 70% data for training and 30% data for testing of the Random Forest model which is 99.69% which is highlighted in the table also. Then with that combination the crop is predicted. This would be very beneficial for farmers for growing their crop without any risk of failure. It also helps in increasing the economy of our country.

6. Conclusion and Future Scope

The proposed system takes the input parameters into consideration and determines the best suitable model at best ratio of training and testing for production of crop. So, the system examine which crop is to be produced in that particular field. This model ensures the farmer to decide on their own that which crop provides them the maximum profit through their agricultural field. By applying machine learning to sensor data, farm management systems are evolving into real time artificial intelligence enabled programs that provide rich recommendations and insights for farmer decision support and action9). Our future work focusses on modifying this model with some other soil characteristics and with more no. of data set.

Farmers decisions, such as land preparation, irrigation, sowing date or fertilizer applications, have also a great influence on crop yield7).

Normally, this prediction is carried out according to the farmers' long-term experience for specific fields, crops and climate conditions. Future agricultural duties may also be carried out by some autonomous farm vehicles that can give high tractive effort at modest speeds. These tractors can detect their location, control the speed, and avoid obstacles while carrying out various agricultural duties including tillage thanks to GPS technology. Numerous agricultural trade exhibitions and agricultural fairs are held in various nations to promote new technologies, and these events allow investors and other technology-based businesses to present their concepts and cutting-edge innovations to the general audience.

References

1) Gholap, J., Ingole, A., Gohil, J., Gargade, S. and Attar, V., 2012. "Soil data analysis using classification techniques and soil attribute prediction" . *arXiv preprint arXiv:1206.1557.*
2) Doshi, Z., Nadkarni, S., Agrawal, R., & Shah, N. (2018, August). "Agroconsultant: Intelligent crop recommendation system using machine learning algorithms". *In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6). IEEE.*
3) Kumar, Avinash, Sobhangi Sarkar, and Chittaranjan Pradhan. "Recommendation system for crop identification and pest control technique in agriculture." *In 2019 International Conference on Communication and Signal Processing (ICCSP), pp. 0185-0189. IEEE, 2019.*
4) Kulkarni, N.H., Srinivasan, G.N., Sagar, B.M. and Cauvery, N.K., 2018, December. "Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique" . In 2018 *3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS) (pp. 114-119). IEEE.*
5) Lokesh.K, Shakti.J, S.Wilson, Tharini.M.S, “Automated crop prediction based on efficient soil nutrient estimation using sensor network”, *July2016,National Conference on Product Design (NCPD 2016)*
6) Pudumalar, S., Ramanujam, E., Rajashree, R.H., Kavya, C., Kiruthika, T. and Nisha, J., 2017, January. "Crop recommendation system for precision agriculture. In 2016 Eighth International Conference on Advanced Computing (ICoAC)" *(pp. 32-36). IEEE.*
7) Thomas, K.T., Varsha, S., Saji, M.M., Varghese, L. and Thomas, E.J., 2020. "Crop Prediction Using Machine Learning. International Journal of Future Generation Communication and Networking," *13(3), pp.1896-1901.*
8) Vijayalakshmi, R., M. Thangamani, M. Ganthimathi, M. Ranjitha, and P. Malarkodi. "An automatic procedure for crop mapping using agricultural monitoring." In *Journal of Physics: Conference Series*, vol. 1950, no. 1, p. 012053. IOP Publishing, 2021.
9) Liakos, K.G., Busato, P., Moshou, D., Pearson, S. and Bochtis, D., 2018. Machine learning in agriculture: A review. *Sensors*, *18*(8), p.2674.
10) Chen, Xiangtuo, and Paul-Henry Cournéde. "Model-driven and data-driven approaches for crop yield prediction: analysis and comparison." *International Journal of Mathematical and Computational Sciences* 11, no. 7 (2018): 334-342.
11) Hanif, Shazia, Muhammad Sultan, Takahiko Miyazaki, and Shigeru Koyama. "Steady-state investigation of desiccant drying system for agricultural applications." *Evergreen* 5, no. 1 (2018): 33-42. https://doi.org/10.5109/1929728
12) Sungkar, Meizka, Tania Surya Utami, Rita Arbianti, and Heri Hermansyah. "The Production of Bioinsecticide Based From Pong-Pong Fruit Seed

Extract by Ultrasonic Waved Extraction Using NADES Solvent." *Evergreen* 7, no. 2 (2020): 303-308. https://doi.org/10.5109/4055237

13) Nugraha, Achmad T., Gunawan Prayitno, Abdul W. Hasyim, and Fauzan Roziqin. "Social Capital, Collective Action, and the Development of Agritourism for Sustainable Agriculture in Rural Indonesia." (2021): 1-12.

14) Sultan, Muhammad, Ibrahim I. El-Sharkawl, and Takabiko Miyazaki. "Experimental study on carbon based adsorbents for greenhouse dehumidification." *Evergr. Jt. J. Nov. Carbon Resour. Sci. Green Asia Strat* 1 (2014).

15) Budihardjo, Mochamad Arief, Nany Yuliastuti, and Bimastyaji Surya Ramadan. "Assessment of Greenhouse Gases Emission from Integrated Solid Waste Management in Semarang City, Central Java, Indonesia." (2021): 23-35.

16) Ridassepri, Arikasuci Fitonna, Fitria Rahmawati, Kinkind Raras Heliani, Jin Miyawaki, and Agung Tri Wijayanta. "Activated carbon from bagasse and its application for water vapor adsorption." *Evergr. Jt. J. Nov. Carbon Resour. Green Asia Strateg* 7 (2020): 409-416.

17) Fageria, Nand Kumar, Virupax C. Baligar, and Ralph Clark. *Physiology of crop production*. crc Press, 2006.

18) Gregory, Peter J., and Timothy S. George. "Feeding nine billion: the challenge to sustainable crop production." *Journal of experimental botany* 62, no. 15 (2011): 5233-5239.

19) Campbell, C.A., Myers, R.J.K. and Curtin, D., 1995. Managing nitrogen for sustainable crop production. *Fertilizer research*, *42*(1), pp.277-296.

20) Battese, G.E., Harter, R.M. and Fuller, W.A., 1988. An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), pp.28-36.

21) Nagendra Kumar Maurya, Vikas Rastogi, and Pushpendra Singh, "Experimental and Computational Investigation on Mechanical Properties of Reinforced Additive Manufactured Component", EVERGREEN Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy, 6 (3) 207-214 (2019). https://doi.org/10.5109/2349296

22) Ang Li, Azhar Bin Ismail, Kyaw Thu, Muhammad Wakil Shahzad, Kim Choon Ng, and Bidyut Baran Saha, "Formulation of Water Equilibrium Uptakes on Silica Gel and Ferroaluminophosphate Zeolite for Adsorption Cooling and Desalination Applications", Evergreen, 1 (2) 37-45 (2014). https://doi.org/10.5109/1495162

23) Jabir Al Salami, Changhong Hu, and Kazuaki Hanada, 'A Study on Smoothed Particle Hydrodynamics for Liquid Metal Flow Simulation' (Kyushu University, 2019).

24) Matheus Randy Prabowo, Almira Praza Rachmadian, Nur Fatiha Ghazalli, and Hendrik O Lintang, "Chemosensor of Gold (I) 4-(3, 5-Dimethoxybenzyl)-3, 5-Dimethyl Pyrazolate Complex for Quantification of Ethanol in Aqueous Solution", Evergreen, 7 (3) 404-408 (2020). https://doi.org/10.5109/4068620

25) Jain, Ankit, Cheruku Sandesh Kumar, and Yogesh Shrivastava. "Fabrication and Machining of Metal Matrix Composite Using Electric Discharge Machining: A Short Review." Evergreen, 8 (4) 740-749 (2021) https://doi.org/10.5109/4742117