

3. METHODOLOGY

Data is a very important part of any Machine Learning Model. Hence, DigiFarm is a user-friendly website designed in such a way that anyone can use it to predict the best crop that can be grown on their soil. To predict the crop we have used Machine Learning (ML) and Artificial Intelligence (AI) technologies. The prediction model is the result of testing the dataset with the best ML algorithms Random forest (RF) classifier and Gradient Boosting (GB) classifier algorithms.

Following are the steps that we have followed to create the crop prediction model:

- 1) Data collection
- 2) Data preparation and analysis
- 3) Choosing the algorithm for training the dataset
- 4) Testing the Machine Learning (ML) model and evaluation
- 5) Deploying

1) Data collection

To train this prediction model we are using the dataset which we have procured from Kaggle website [19] which is shown in Figure 1. This dataset was built by augmenting datasets of rainfall, climate and fertilizer data available for India which was gathered over the period by Indian Chamber of Food and Agriculture (ICFA), India.

	A	B	C	D	E	F	G	H
1	N	P	K	temperature	humidity	ph	rainfall	label
2	90	42	43	20.87974371	82.00274423	6.502985292	202.9355362	rice
3	85	58	41	21.77046169	80.31964408	7.038096361	226.6555374	rice
4	83	95	50	26.51682337	77.79913575	5.50947065	108.8547508	banana
5	119	90	48	28.66725136	79.59242542	5.986442306	118.2583441	banana
6	78	42	42	20.13017482	81.60487287	7.628472891	262.7173405	rice
7	69	37	42	23.05804872	83.37011772	7.073453503	251.0549998	rice
8	38	15	30	28.91862016	48.13974548	5.075504537	97.01331604	mango
9	12	37	30	31.09779147	47.41196659	4.546466109	90.28624348	mango
10	38	19	31	34.73823882	49.08864345	5.855119268	90.65022183	mango
11	8	33	29	29.98080499	49.48613279	6.442393461	91.82271568	mango

Figure 2: Dataset used to train Machine Learning Model

2) Data preparation and analysis

The Dataset which is shown in Figure 1 has 2201 samples among which we have used 90% (i.e. 1980 samples) for the purpose of training. The remaining 10% (i.e. 221 samples) are used for testing purposes. The dataset contains 8 attributes (they are Nitrogen, Phosphorous, Potassium,

Temperature and humidity of the region, pH of the soil, Rainfall in mm in that region and the crop name. The dataset contains 22 distinct categories (i.e., Apple, Banana, Blackgram, Chickpea, Coffee, Cotton, Grapes, Jute, Kidneybeans, Lentil, Maize, Mango, Mothbeans, Mungbean, Muskmelon, Orange, Papaya, Pigeonpeas, Pomegranate, Watermelon, Rice and Coconut).

Before training our dataset we conducted the following preliminary analysis of the dataset:

- i) Finding out some statistical information about the data which is summarized in the Table 1

Statistical Parameters	Nitrogen	Phosphorous	Potassium	Temperature	Humidity	ph	Rainfall
Count	2200.0	2200.0	2200.0	2200.0	2200.0	2200.0	2200.0
Mean	50.55	53.36	48.14	25.61	71.48	6.48	103.46
Standard Deviation	36.91	32.98	50.64	5.063	22.26	0.77	54.95
Minimum Value	0.0	5.0	5.0	8.82	14.25	3.50	20.21
25%	21.0	28.0	20.0	22.76	60.26	5.97	64.55
50%	37.0	51.0	32.0	25.59	80.47	6.42	94.86
75%	84.25	68.00	49.0	28.56	89.94	6.92	124.26
Maximum Value	140.0	145.0	205.0	43.67	99.98	9.93	298.56

Table 1: Basic statistical information about the dataset

ii) We found out the correlation between different attributes which is summarized in the following Table 2.

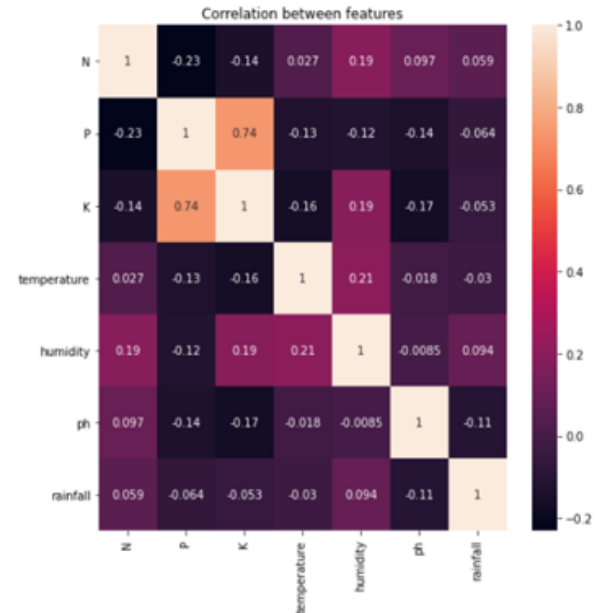


Table 2: Correlation between features

iii) Since there were no missing cells or null values we then moved to the next step i.e., choosing the algorithm and training it.

3) Choosing the algorithm for training the dataset

In this research work, we have considered “crop name” as target label and used “Multiclass Classification”. We have used Gradient Boosting Classifier and Random forest algorithms for training the ML model.

Gradient Boosting (GB) Classifier Algorithm

Gradient boosting is a Machine Learning technique for regression, classification and other tasks which was invented by Leo Breiman. It is a boosted ensemble of tree as opposed to a bagged ensemble they have very low interpretability because the second tree in the model no longer predicts the same target as the original model and the subsequent trees in the model seek to predict how far off the original predictions were from the truth by using the residuals from the prior trees. In this way each subsequent tree in the GB model slowly reduces the overall error of the previous trees. This enables GB models to have very high predictive power but low interpretability. In addition GB models are quite prone to overfitting the training data to combat there are several extra hyper parameters that are not needed in forests. They are learning rate which controls how you add subsequent trees together and also regularisation in the form of ridge and lasso hyper parameters.

Random Forest (RF) Classifier Algorithm

The Random decision forests or Random forests are an ensemble method (it groups multiple Decision tree predictors) used for regression, classification and other ML tasks which was developed by Leo Breiman. In this each tree makes their own prediction and they are aggregated into a final prediction either by votes for classification problems or as an average for regression problems. If all of the decision trees are the same then each tree will predict the same output this is where the random part of the random forests come into play. There are two aspects of randomness involved. First is what features in each decision tree a random subset of features is chosen. Second aspect of randomness is using only a sample of the training data each time a tree is fit. The idea is to have each row and feature utilized in at least one of the decision trees, however not to use all the features in any single decision tree. This allows us to build trees that are not correlated while adding variation to our collection of models and reduce the risk of overfitting. Even with this approach overfitting is still a concern that can be solved by setting the max depth. Max depth is the number of questions asked before we reach the prediction. We limit the depth to reduce the risk of overfitting.

4) Testing the Machine Learning (ML) model and evaluation

On training the GB classifier algorithm and RF classifier algorithm models by the dataset, the accuracy that we got from GB algorithm was 0.996 and that of RF was 0.998. On testing both the models with 10% samples of the dataset we found that the accuracy that we got from the GB Classifier algorithm is 0.982 and that of the RF classifier was 0.989. Since the accuracy of RF was high in both training and testing phases we chose RF model to deploy and use it for predicting the crop.

5) Deploying

For deploying this ML model we have used IBM cloud services. Since it is deployed in a cloud, we are using the ML Model for prediction through Application Program Interface (API).

3. RESULT AND DISCUSSION

In this section we shall see the outcome of our platform “DigiFarm” which is designed to carefully and accurately predict the most suitable crop that the farmer can produce in his region. DigiFarm is designed to equip the farmers with digitized farming so that they can make the most out of their crops. With the aid of this platform, they can receive precise information about which crops would be most suitable for their land. User can predict their crop based on two methods:

Method 1: By making use of region's weather conditions, pH value of the soil, rainfall pattern and soil composition (i.e., nitrogen, phosphorus and potassium) as inputs

Method 2: By making use of place/location and current season as inputs

3.1 Software Compatibility

Our platform DigiFarm is compatible with latest versions of browsers such as Google Chrome, Microsoft Edge, Mozilla Firefox etc. The Front-end of DigiFarm is designed using HTML 5, CSS 3 and JavaScript whereas for the Back-end we have used Django 3.2.5 (i.e., a Python back-end framework).

3.2 Home Page



Figure 1: Home Page of the DigiFarm

DigiFarm platform along with its name is shown in Figure 1. The logo displayed on the top left corner has two central elements: a hand and the water. The water is representative of rivers and oceans that forms the backbone of the irrigation system in Indian agriculture. The hand that is holding a plant represents the farmers who grow the crops. Since India is an agrarian economy,

not only is the population dependent on the farmers for food, but the national economy is also dependent on the yield from the primary sector.

This page connects the user to all the different pages on the platform. The Navigation bar on the homepage as shown in Figure 1 has different buttons for various purposes such as ChatBot, Prediction and News. Additionally, there are two buttons at the center of the homepage: one to get the detailed tutorial on how to use the website and make the best use of it for the users and another button is to predict the most suited crop(s).

Now, let us see what each section of the platform does in detail:

3.3 Prediction Section

This part of DigiFarm is the main part of the platform. The “Prediction” section as shown in Figure 2 can be used to predict the most suitable crops that can be sown on their land by using two different methods.

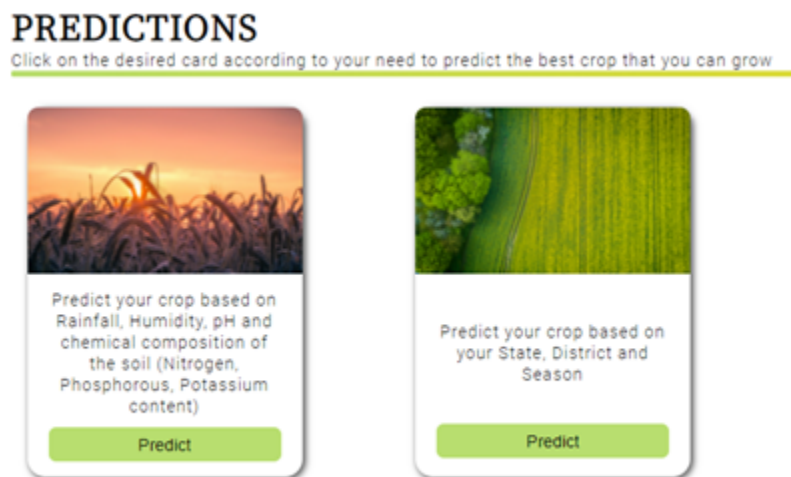


Figure 2: Predictions Section (i.e. Different methods for predicting the crop)

3.3.1 Method 1

As shown in the Figure 3(a) users can predict the crop by entering details such as rainfall, humidity, and temperature, and pH value along with the composition of the soil (i.e. nitrogen, phosphorus, and potassium content), The system will predict the crop which is most suitable for the given geographical conditions.

Instructions

In order to predict the best crop, Enter the values for the following fields. You have to get Nitrogen, Potassium, Phosphorous and pH values of the soil by contacting your nearest Government Soil Testing Laboratory. Incase if your facing any difficulties please free to open our ChatBot by clicking the button at the bottom of the screen and get your problem solved.

NOTE: Fields marked with (*) are compulsory.

Nitrogen Content*	Location*
<input type="text" value="N"/>	<input type="text" value="Enter yor place name. Example: bengaluru"/>
Phosphorous Content*	Temperature in °C
<input type="text" value="P"/>	<input type="text" value="Example: 25"/>
Potassium Content*	Humidity
<input type="text" value="K"/>	<input type="text" value="Example: 60"/>
pH of the soil*	<input type="button" value="Submit"/>
<input type="text" value="Example: 7"/>	
Rainfall in mm*	
<input type="text" value="Example: 900"/>	

Figure 3(a): Prediction of crops using method 1

Consider an instance, when the user enters the values of nitrogen content as 20, phosphorus content as 89, potassium content as 40, pH value as 6, rainfall as 700, location as Bengaluru the Machine Learning (ML) model will predict the crop based on the values entered and the result is displayed as Coconut as shown in the Figure 3(b).

Result

The best crop that can be sown on your soil is **Coconut**

Figure 3(b): The output based on method 1

Similarly, we can get different crop names as output (as shown in the figures 3(c) to 3(f)) based on the different input combinations.

Figure 3(c): For the input values $N=89$, $P=58$, $K=38$, $Temp=23^{\circ}C$, $Humidity=83\%$, $pH=6.3$ And $Rainfall=221mm$	Result <hr/> The best crop that can be sown on your soil is Coconut
Figure 3(d): For the input values $N=86$, $P=76$, $K=54$, $Temp=29^{\circ}C$, $Humidity=80\%$, $pH=5.9$ And $Rainfall=90mm$.	Result <hr/> The best crop that can be sown on your soil is RICE
Figure 3(e): For the input values $N=36$, $P=125$, $K=196$, $Temp=37^{\circ}C$, $Humidity=80\%$, $pH=6.1$ And $Rainfall=66mm$	Result <hr/> The best crop that can be sown on your soil is Banana
Figure 3(f): For the input values $N=34$, $P=140$, $K=198$, $Temp=21^{\circ}C$, $Humidity=93\%$, $pH=5.75$ And $Rainfall=115mm$.	Result <hr/> The best crop that can be sown on your soil is Grapes

3.2.1.1 Discussion on prediction of crop (method 1) dataset

As discussed in the section 2 i.e methodology We have used Gradient Boost Classifier and Random Forest classifier algorithms for training. A confusion matrix is a table that is used to describe the performance of a classification model on a set of test data for which the true values are known. It represents the ways in which your classification model is confused when it makes predictions. The confusion matrix that was generated on running the test data with Gradient Boost model and Random Forest model is as shown in the Figure x and y respectively where the numberings 0-21 refers to the crops as shown in the lookup table 4. For instance the Gradient Boosting Model is confused between crop 8 i.e. kidneybeans and crop 20 i.e. rice 7 times (in other words this model wrongly predicted the crop 8 as crop 20, 7 times).

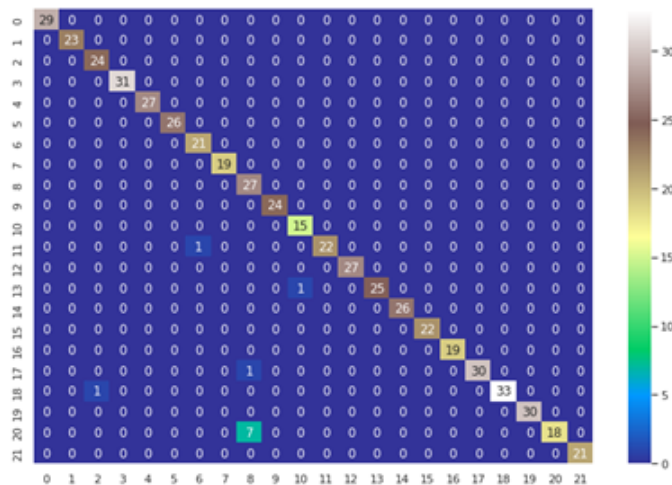


Figure x: Confusion Matrix for GB Model

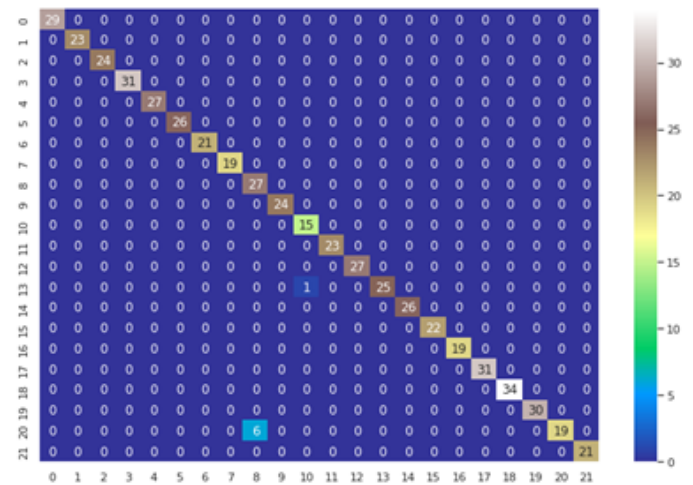


Figure y: Confusion Matrix for RF Model

Index	Corresponding Crop	Index	Corresponding Crop	Index	Corresponding Crop
0	Apple	8	Kidneybeans	16	Papaya
1	Banana	9	Lentil	17	Pigeonpea
2	Blackgram	10	Maize	18	Pomegranate
3	Chickpea	11	Mango	19	Watermelon
4	Coffee	12	Mothbeans	20	Rice
5	Cotton	13	Mungbean	21	Coconut
6	Grapes	14	Muskmelon		
7	Jute	15	Orange		

Table 4: Crop lookup table

A **true positive** is an outcome where the model *correctly* predicts the *positive* class. Similarly, a **true negative** is an outcome where the model *correctly* predicts the *negative* class. A **false positive** is an outcome where the model *incorrectly* predicts the *positive* class. And a **false negative** is an outcome where the model *incorrectly* predicts the *negative* class. [3]

Table 2 summarizes the true positive rate, true negative rate, false positive rate and false negative rate for Random Forest model and Gradient boosting model.

	Random Forest Model	Gradient Boosting Model
True positive rate	0.990	0.981
True negative rate	0.990	0.999
False positive rate	0.0004	0.0008
False negative rate	0.009	0.018

Table 2: What name can i give for this?

The way we evaluate ML models is generally through the parameters accuracy, precision, recall and F1 score.

Accuracy is the ratio of total number of correctly predicted data points to the total number of all the data points i.e.,

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / (\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives})$$

Precision is defined as the ratio of True Positives to the sum of True positives and false positives i.e.,

$$\text{Precision} = (\text{True Positives}) / (\text{True Positives} + \text{False Positives})$$

Recall is defined as the ratio of True Positives to the sum of True positives and false negatives i.e.,

$$\text{Precision} = (\text{True Positives}) / (\text{True Positives} + \text{False Negatives})$$

F1 score is a simple way to compare two classifiers and is defined as the harmonic mean of recall and precision.

$$\text{F1 score} = 2 / ((1/\text{Precision}) + 1/\text{Recall})$$

$$= (\text{True Positives}) / (\text{True Positives} + ((\text{False Positives} + \text{False Negatives})/2))$$

Table 3 summarizes the values for these parameters that we got on testing Random Forest and Gradient Boosting models.

Performance Measures	Random Forest Model	Gradient Boosting Model
Accuracy	0.989	0.982
Recall	0.990	0.981

Precision	0.990	0.981
F1 Score	0.990	0.981

Table 3: Performance measures (i.e. Accuracy, Recall, Precision, F1 Score) of predictive models

As we can see in the Table 3 RF Model outperformed Gradient Boosting Model in all aspects so we chose RF model to deploy and use it for predicting the crop. This system of crop prediction results in accuracy and efficiency which is unprecedented.

3.3.2 Method 2

Another option for predicting the crops is by mentioning their state, district, and season as shown in Figure 5(a).

Instructions

In order to predict the best crop, select your State, District and Season from the dropdown box.

NOTE: Fields marked with (*) are compulsory.

State*
Karnataka ▼

District*
BANGALORE RURAL ▼

Season*
Kharif ▼

Submit

Figure 5(a): Input fields for approach 2

For example:- When we enter the state as Karnataka, district as Bengaluru Rural, season as kharif (ref Figure 5(a)), we get the crop prediction of ONION, DRY GINGER, RAGI, BAJRA, MAIZE, RICE, GRAM and many more as shown in Figure 5(b).

Result

The best crops that can be according to the state, district and season specified by you are:

- Niger seed
- Dry ginger
- Groundnut
- Onion
- Rapeseed & Mustard
- Soyabean
- Urad
- Gram
- Cowpea(Lobia)
- Moong(Green Gram)
- Small millets
- Sesamum
- Bajra
- Other Kharif pulses
- Potato
- Paddy
- Ragi
- Castor seed
- Dry chillies
- Rice
- Peas & beans (Pulses)
- Cotton(lint)

Figure 5(b): Sample result from the second method for Crop Prediction

3.3.2.1 Discussion on prediction of crop (method 2) dataset

For this second method we have made use of a dataset which is different from the first one. This dataset contains the name of the crops which gave the best yield in each seasons (i.e. Autumn, Kharif, Rabi, Summer, Winter). This crop and season data is available for all the districts of India. This dataset has 4 columns (as shown in Figure 6) namely the State Name, District Name, Cropping Season and crop names.

For predicting the crops in this method we are not using any ML model instead we are searching the dataset and displaying the crop names according to the information (i.e. State name, District name and the preferred season) entered by the user.

	A	B	C	D
1	State_Name	District_Name	Season	Crop
2	Andaman and Nicobar Islands	NICOBARS	Kharif	Arecanut
3	Uttarakhand	PAURI GARHWAL	Rabi	Rapeseed & Mustard
4	Uttarakhand	PAURI GARHWAL	Rabi	Wheat
5	Tamil Nadu	MADURAI	Whole Year	Cashewnut
6	Tamil Nadu	MADURAI	Whole Year	Coconut
7	Tamil Nadu	MADURAI	Whole Year	Coriander
8	Karnataka	KOPPAL	Kharif	Sesamum
9	Karnataka	KOPPAL	Kharif	Small millets
10	Karnataka	KOPPAL	Kharif	Sunflower

Figure 6: Crop prediction dataset for method 2

Farmers generally wish to continue growing the same crops on their land either to avoid risk or due to lack of awareness. However, our platform would enable them to go beyond their regular pattern by providing accurate and precise information about the crop which they can sow on their land to get the maximum yield.

3.4 AgriBot

In the current growing age of digitization, Artificial Intelligence (AI) powered chatbots are playing a leading role by exemplifying the function of a virtual assistant that could manage a conversation via speech or textual methods. It makes use of voice or textual queries to get answers, perform actions and recommendations according to user needs. They are adaptable to the user's individual language usages, searches, and preferences with continuing use. With the advent of AI, virtual assistants can be seen penetrating to the nook and corner of the world [16]. A conversational bot with a voice and/or chat interface can play a principal role in solving the user queries by giving instant service. This drastically reduces waiting time for the users (i.e. mainly farmers) to get their queries solved. The ChatBot as shown in Figure 4 is an AI based bot called

the “AgriBot”. It is created using the IBM Watson Chatbot services (i.e. a part of IBM Cloud services).

The result of the AgriBot is the easy accessibility of DigiFarmplatform for the users. The AgriBot provides user three options (i.e. as shown in the Figure 7) which can be availed one at a time. This helps the user to get their queries solved.

The different options available for users are:

- 1) To guide the users about the platform and how to make use of it.
- 2) Various methods to predict the crop
- 3) Directs users to a Google Form which they can fill for any further queries

The presence of a AgriBot on this platform simplifies the user experience which is the broader aim of the Digifarm.

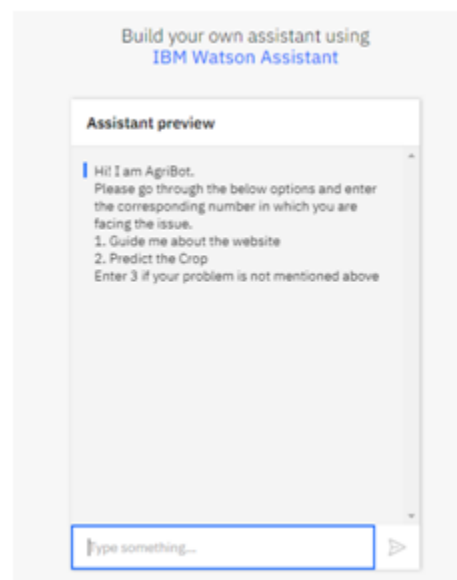


Figure 7: AgriBot

3.6 News Segment

The news segment results in bringing together the latest information about agriculture from across the globe as shown in Figure 7. The idea behind this is to keep our users about the technological developments happening in the agricultural field from different parts of the world. This enables them to learn from these techniques and apply the suitable ones on their land. The result would be increased productivity, developing a nature of taking risks and also equipping them with the most updated advancements which they may find suitable for their farm.

In the news section of DigiFarm, we have used RSS (Really Simple Syndication) news feed from “The Hindu Agri-Business” section.

NEWS

Find out the latest NEWS about agriculture below



The Hindu - Agri-Business

[India's agriculture exports can be doubled in 'medium term', says ITC Chairman](#)
Thu, 30 Sep 2021 17:49:08 +0530
"There is an urgent need to deal with the agriculture value chain, which would help in improving the quality of income in rural areas and ushering in socio-economic transformation in the country, ITC Chairman Sanjiv Puri said.

[Data / Most farmers sold to private traders in 2019, new survey data shows](#)
Wed, 15 Sep 2021 10:44:35 +0530
Data also show that a majority of the farmers were unaware of the Minimum Support Price regime

[Cabinet decision to hike MSP for Rabi crops will ensure max. remunerative price for farmers, PM](#)
Thu, 09 Sep 2021 08:08:00 +0530
Prime Minister Narendra Modi on Wednesday said the Cabinet's approval to increase the Minimum Support Prices (MSP) for Rabi crops will ensure maximum

[Amid protests, Centre hikes minimum support price for rabi crops](#)
Thu, 09 Sep 2021 09:10:27 +0530

Courtesy: [THE HINDU](#)

Feed Widget
Activate Windows
Go to Settings to activate Windows

Figure 7: News section

The idea is to build a platform which solves all the queries of the farmers holistically and comprehensively. The main aim of DigiFarm is to help the farmers in increasing their production and yield per square by choosing the right crop for their field at the right time. Also, it enables them to sell their crop for maximum revenue. DigiFarm would result in digitization of agriculture on a global scale. However, it would prove extremely beneficial to farmers in India who find it difficult to access the information as their outreach is confined which results in using obsolete methods despite the advancements. DigiFarm, with the aid of technology will help them with the best possible information for their farm using a single platform.

5. References

- [1] Patel, Sanoj Kumar, Anil Sharma, and Gopal Shankar Singh. "Traditional agricultural practices in India: an approach for environmental sustainability and food security." *Energy, Ecology and Environment* 5.4 (2020): 253-271.
- [2] FAO (2016) The State of Food and Agriculture, Climate Change, Agriculture and Food Security. Food and Agriculture Organization of the United Nations Rome, 2016. www.fao.org
- [3] Adams MW, Ellingboe AH, Rossman EC (1971) Biological uniformity and disease epidemics. *Bioscience* 21(21):1067–1070
- [4] Sofia, P. K., Rajendra Prasad, and V. K. Vijay. "Organic farming-tradition reinvented." (2006).
- [5] Singh GS, Ram SC, Kuniyal JC (1997) Changing traditional land use patterns in the Great Himalayas: a case study of Lahaul Valley. *J Environ Syst* 25:195–211
- [6] Jeeva SRDN, Laloo RC, Mishra BP (2006) Traditional agricultural practices in Meghalaya, North East India. *Indian J Trad Knowl* 5(1):7–18
- [7] Pradhan, Aliza, et al. "Potential of conservation agriculture (CA) for climate change adaptation and food security under rainfed uplands of India: A transdisciplinary approach." *Agricultural Systems* 163 (2018): 27-35.
- [8] Lincoln NK (2019) Learning from indigenous agriculture. *Nat Sustain* 2(3):167
- [9] Johns, Timothy, Bronwen Powell, Patrick Maundu, and Pablo B. Eyzaguirre. "Agricultural biodiversity as a link between traditional food systems and contemporary development, social integrity and ecological health." *Journal of the Science of Food and Agriculture* 93, no. 14 (2013): 3433-3442.
- [10] Rana, Ram Bahadur, Chris Garforth, Bhuwon Sthapit, and Devra Jarvis. "Influence of socio-economic and cultural factors in rice varietal diversity management on-farm in Nepal." *Agriculture and human values* 24, no. 4 (2007): 461-472.
- [11] Gruyere G, Mehta-Bhatt P, Sengupta D. Bt cotton and farmer suicides in India: reviewing the evidence.
- [12] Nagaraj K. Farmers' suicides in India: magnitudes, trends and spatial patterns.

[13] Kennedy, Jonathan, and Lawrence King. "The political economy of farmers' suicides in India: indebted cash-crop farmers with marginal landholdings explain state-level variation in suicide rates." *Globalization and health* 10.1 (2014): 1-9.

[14] R Shamshiri, R., Weltzien, C., Hameed, I. A., J Yule, I., E Grift, T., Balasundram, S. K., ... & Chowdhary, G. (2018). Research and development in agricultural robotics: A perspective of digital farming.

[15] Suresh, G., Kumar, A. S., Lekashri, S., & Manikandan, R. (2021). Efficient crop yield recommendation system using machine learning for digital farming. *International Journal of Modern Agriculture*, 10(1), 906-914.

[16] U. Bharti, D. Bajaj, H. Batra, S. Lalit, S. Lalit and A. Gangwani, "Medbot: Conversational Artificial Intelligence Powered Chatbot for Delivering Tele-Health after COVID-19," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 870-875, doi: 10.1109/ICCES48766.2020.9137944.

[17] [https://en.wikipedia.org/wiki/Out-of-bag_error#:~:text=Out%2Dof%2Dbag%20\(OOB,utilizing%20bootstrap%20aggregating%20\(bagging\)](https://en.wikipedia.org/wiki/Out-of-bag_error#:~:text=Out%2Dof%2Dbag%20(OOB,utilizing%20bootstrap%20aggregating%20(bagging))

[18] <https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative>

[19] <https://www.kaggle.com/siddharthss/crop-recommendation-dataset>