International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Generalized Smart Agriculture System with the Prediction of Crop Yield Per Year

Priyanka Mehta[1], Niprojit Ghosh[2], Deepjyoti Chakraborty[3], Oindrila Ganguly[4], Vedatrayee Chakraborty[5]

Department of Electronics and Communication Engineering, B. P. Poddar Institute of Management and Technology, Kolkata, India

Abstract: ***Agriculture is one of the most essential and widely practiced occupations in India and it also plays a vital role in the development of our country [1]. Around 60 percent of the total land in this country is used for agriculture to meet the needs of 1.2 billion people, so improving crop prediction is therefore seen as a significant aspect of agriculture. Basically if one has a piece of land, she or he needs to know what kind of crop can be grown in this area [2, 3]. Agriculture also depends on various soil properties. Production of crops is also a difficult task since it involves factors like soil type, temperature, humidity etc. If it is possible to find the crop before sowing it, it would be of great help to the farmers and the other people involved to make appropriate decisions on the storage and business side [4-6]. The proposed project would solve agricultural problems by monitoring the agricultural area on the basis of soil properties and recommending the most appropriate crops to farmers, thereby helping them to significantly increase the crop productivity. In this work, the modelling makes use of different Machine Learning Techniques such that it gives the information which includes soil ingredients, rainfall, quantity of crops, production of crops with respect to temperature of that area etc. which would definitely be beneficial for the farmers.***

Keywords: Machine Learning, Crop prediction, Rainfall prediction, soil ingredients, Crop recommendation

1. Introduction

Agriculture always plays a major role in the lives of every individual. From the old time itself agriculture is considered to be one of the main practice practiced in India. In older times, people used to cultivate crops in their own lands in order to meet their requirements and interest.

The decision of the farmers regarding which type of crops and vegetables to grow in his land generally depends on his intuition and many other factors such as making huge profits within a short period of time, lack of awareness about the demand in the market and also the soil's potential to support the growth of a particular type of crop and many more.

Agriculture produce is subjected to various risks, which are not only confined to production risk pertaining to weather, pest but also the demand and supply of various countries, other policy and economic factors. High price volatility has been a major concern in the past few years both for farmers and consumers. The main purpose of price prediction is to help producers manage their price risk and take informed decisions. Machine Learning has proved to be better than the traditional time series method of price prediction using many linear and non-linear forecasting models.

The scope of this work is to determine the crop yield of an area by considering dataset with some features which are important or related to crop production such as temperature, moisture, rainfall, and production of the crop in previous years. To predict a continuous value, regression models are used. It is a supervised technique. The coefficients are pre-processed and fit into the trained data during training and construction the regression model. The main focus here isto be able to produce the final output of the district-wise crop yield production per year. As a result of which the farmers would be benefited to make their decisions wisely without being subjected to any losses.

2. Theory

Agriculture is the practice of cultivating plants trees and live stocks. Agriculture is and was the key development of the rise of sedentary human civilization, whereby farming of domesticated species created food surpluses that enabled people to live in cities and villages.

After gathering wild grains beginning at least 100, 000 years ago, farmers began to plant them around 11, 000 years ago. Pigs, cows, and cattle were domesticated over 10, 000 years ago. Plants were independently cultivated in at least 11 regions of the world. Industrial agriculture based on large-scale monoculture in the twentieth century came to dominate agricultural output, though about 2 billion people still depended on agriculture [7, 8].

Modern agronomy, plant breeding, agrochemicals such as pesticides and fertilizers, and technological developments have sharply increased crop yields, while causing widespread ecological and environmental damage. Selective breeding and modern practices in animal husbandry have similarly increased the output of meat, but have raised concerns about animal welfare and environmental damage. Environmental issues include contributions to global warming, depletion of aquifers, deforestation, antibiotic resistance, and growth hormones in industrial meat production. Agriculture is both a cause of and sensitive to environmental degradation, such as biodiversity loss, desertification, soil degradation and global warming, all of which can cause decreases in crop yield. Genetically modified organisms are widely used, although some are banned in certain countries.

To advantage the cultivating from the new worldwide market, get admission to potential outcomes, the inward rural promoting device inside the United States of America moreover wishes to be joined and strengthened. In

interesting, the commercial centre contraption must be revived to:

Provide impetuses to Farmer to deliver more. Pass on the changing over wishes of the purchasers to the makers to empower producing making arrangements. Foster genuine challenge a considerable lot of the market players and To improve the offer of Farmers in the last expense of his rural produce.

Today the farmers develop crops dependent on the experience picked up from the past age. Since the customary technique for cultivating is polished there exists an overabundance or shortage of yields without gathering the real necessity. The farmers don't know about the interest that happens in the current horticultural economy. This results in the misfortune to the Farmers.

3. Proposed Model

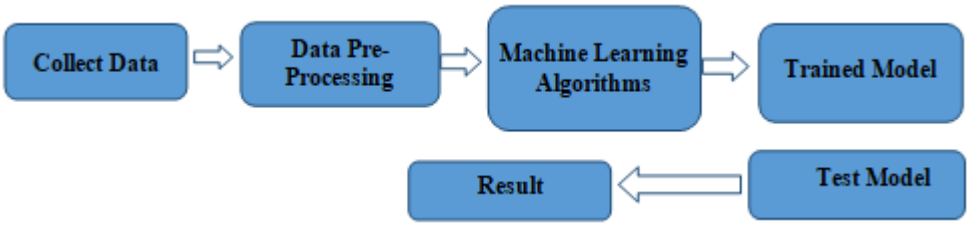

Figure: Model of the System

The proposed project is a model which can predict the crop based on the soil nutrient values given as the input [9, 10]. We have taken a diverse dataset concentrating on various factors like Production, Area, Temperature, Rainfall, Air Quality Index, Amount of NO_2, SO_2, Suspended particular matter since studies show that their excessive presence can have grave effects on photosynthesis (excessive SO2 causes the widening of stomata and leads to excessive water loss, excessive NO2 and SO2 causes acid rain which can have acute effects on the growth of crops) leading to poor and unhealthy crop yield.

1) Collection of dataset

The accuracy of a machine learning algorithm may depend on the number of parameters used and to the extent of correctness of the dataset [6]. Our dataset contains the N, P, K, and pH values of different kinds of soils as attributes and it also contains the corresponding crops that can be grown in that soil as label. Thus, by using an appropriate machine learning algorithm we can train the dataset to predict the most suitable crop that can be grown under the given input parameters.

2) Data pre-processing

Data preprocessing is the second step and it contains two steps. Original dataset can contain lots of missing values so initially all these should be removed. Missing values are denoted by a dot in the dataset and their presence can deteriorate the value of entire data and it can reduce the performance. So, to solve this problem we replace these values with large negative values which will be treated as outliers by the model. Generating the class labels is the second step. Since we are using a supervised learning method, for each entry in the dataset there should be a class label which is created during the preprocessing step.

3) Machine Learning Algorithms

Different Machine Learning algorithms have been used in order to fetch the desired results. They are as follows:

a) k-Nearest Neighbour:

In this algorithm, the input provided will be the k nearest training examples of the dataset and the output will depend on whether it is a classification or regression problem. Basically, it works based on the minimum distance from the given input value which is soil values to the trained values to find the nearest k neighbours and afterwards those with majority is taken to be as the output prediction to predict the crop label. To find which is most similar to the given instance distance measure is used. Mostly, by default Euclidean distance is used as a distance measure. It is calculated by the given formula [11],

$$d = \sqrt{[(x2 - x1)^2 + (y2 - y1)^2]}$$

where,

- (x1, y1) are the coordinates of one point.
- (x2, y2) are the coordinates of the other point.
- d is the distance between (x1, y1) and (x2, y2).

b) Decision Tree:

A decision tree is a non-parametric method of supervised learning technique. Throughout the process a tree like structure is formed. In this, the dataset is broken down to build upon the tree subsequently. Finally, the resulting output is a tree like structure with both decision nodes and leaf nodes. Decision nodes can either have two or more branches while the leaf nodes indicate the final nodes representing classification or regression result. The topmost node is the root node and the one with higher gain (or gini index) value is taken to be the root. Decision trees have the ability to classify both categorical and numerical data.

Regression Analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (predictor). The Regression model used for prediction process are as follows:

- Gradient Boosting Regressor
- XGBoost Regressor

Gradient Boosting Regressor: It is a Machine Learning Technique for regression, classification tasks, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. The key idea is to set the target outcomes for the next model in order to minimize the error.

The following steps are involved in gradient boosting:

- F0 (x) – with which we initialize the boosting algorithm – is to be defined:

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$$

- The gradient of the loss function is computed iteratively:

$$r_{im} = -\alpha \left[\frac{\partial (L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)},$$

where α is the learning rate

- Each hm (x) is fit on the gradient obtained at each step
- The multiplicative factor γm for each terminal node is derived and the boosted model Fm (x) is defined:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

4) Trained Model

Trained models are obtained after applying the dataset to the machine learning algorithms. Our project suggests a crop prediction system which is based on the K-Nearest Neighbour algorithm. Soil properties such as amount of nitrogen dioxide, sulphur dioxide, suspended particulate matter are taken as input to the model since their excessive presence can have grave effects on photosynthesis leading to poor and unhealthy crop yield. So in order to inform the farmers about these conditions we are training the datasets of crop yield of all the states according to these factors too.

4. Result and Discussion

We have taken a diverse dataset concentrating on various factors like Production, Area, Temperature, Rainfall, Air Quality Index, etc.

In this project, we have trained all the state wise data of India according to the number of lines of data we have. Also we are training the state wise information followed by different crop details including area, production, rainfall, yield (tones/ hectares).

```
In [4]: df_Rain['Season']=df_Rain['Season'].str.strip()
        d=pd.merge(df_Rain,new_df,on=['State_Name','Crop','Season','Area','Production','Crop_Year'],how='outer')

In [5]: d.drop(['Unnamed: 0'],axis=1)
```

Out[5]:

	State_Name	Crop_Year	Season	Crop	Area	Production	Rainfall	Yield (Tonnes/Hectare)	District_Name
0	Andaman and Nicobar Islands	2000	Kharif	Arecanut	1254.0	2000.0	2763.2	1.594896	NICOBARS
1	Andaman and Nicobar Islands	2000	Kharif	Other Kharif pulses	2.0	1.0	2763.2	0.500000	NICOBARS
2	Andaman and Nicobar Islands	2000	Kharif	Rice	102.0	321.0	2763.2	3.147059	NICOBARS
3	Andaman and Nicobar Islands	2000	Whole Year	Banana	176.0	641.0	2763.2	3.642045	NICOBARS
4	Andaman and Nicobar Islands	2000	Whole Year	Cashewnut	720.0	165.0	2763.2	0.229167	NICOBARS
...	...	...	...	...	...	...	...	...	...
249507	West Bengal	2014	Summer	Rice	306.0	801.0	NaN	2.617647	PURULIA
249508	West Bengal	2014	Summer	Sesamum	627.0	463.0	NaN	0.738437	PURULIA
249509	West Bengal	2014	Whole Year	Sugarcane	324.0	16250.0	NaN	50.154321	PURULIA
249510	West Bengal	2014	Winter	Rice	279151.0	597899.0	NaN	2.141848	PURULIA
249511	West Bengal	2014	Winter	Sesamum	175.0	88.0	NaN	0.502857	PURULIA

249512 rows × 9 columns

5) Testing Data

While training or testing the data, the data set is split into two sets. Here, in this model we have done 70% Training and 30% Testing. With correct implementation of Machine Learning techniques, Feature Engineering, Outlier Correction, and extensive use of multiple ensemble techniques like XGBoost, Gradient Boost we have achieved 90% accuracy in predicting the crop yield.

```
Test MAPE: 29.2098
Train MAPE: 26.2952
Test SMAPE: 23.3741
Train SMAPE: 14.8878
Test R2: 0.8340
Train R2: 0.9656
Test wMAPE: 27.9487
Train wMAPE: 20.1443
Test wSMAPE: 22.4524
Train wSMAPE: 14.1616
Test wR2: 0.8465
C:\anaconda\lib\site-packages\sklearn\utils\validation.py:70: FutureWarning: Pass sample_weight=[2.53542484 2.66963319 3.021547
47 ... 3.82304886 3.45761694 3.7057965 ] as keyword args. From version 1.0 (renaming of 0.25) passing these as positional argum
ents will result in an error
  warnings.warn(f"Pass {args_msg} as keyword args. From version "
C:\anaconda\lib\site-packages\sklearn\utils\validation.py:70: FutureWarning: Pass sample_weight=[2.91648279 2.26271358 3.683554
47 ... 3.39299539 3.880034   3.16419194] as keyword args. From version 1.0 (renaming of 0.25) passing these as positional argum
ents will result in an error
  warnings.warn(f"Pass {args_msg} as keyword args. From version "
Train wR2: 0.9678
```

5. Results

We have trained the results of all the states of India.

Since, Tamil Nadu is the 3rd largest producer of rice crop and their government site had all the updated data year wise, we predicted the final output considering that, so as to get a better understanding and analysis of district wise crop yield production. Below are the results of our model:

```
In [45]: test=test_[continuous+categories_catcode]

In [46]: predicted= pd.concat([test_XX, pd.DataFrame(clf.predict(test),columns=['Pred '])],axis=1)

In [47]: predicted
```

Out[47]:

	Season	Crop_Year	State_Name	District_Name	Max	Mean	Min	Crop	Area	Production	...	so2	no2	rspm	
0	Rabi	2020	TAMIL NADU	ARIYALUR	0.893953	0.520934	-0.737702	Rice	19628	87227	...	84.123502	100.019805	552.942203	615.58
1	Rabi	2020	TAMIL NADU	COIMBATORE	0.893953	0.520934	-0.737702	Rice	669	2284	...	84.123502	100.019805	552.942203	615.58
2	Rabi	2020	TAMIL NADU	CUDDALORE	0.893953	0.520934	-0.737702	Rice	91927	406593	...	84.123502	100.019805	552.942203	615.58
3	Rabi	2020	TAMIL NADU	DHARMAPURI	0.893953	0.520934	-0.737702	Rice	15564	75766	...	84.123502	100.019805	552.942203	615.58
4	Rabi	2020	TAMIL NADU	DINDIGUL	0.893953	0.520934	-0.737702	Rice	5203	29048	...	84.123502	100.019805	552.942203	615.58
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
81	Kharif	2020	TAMIL NADU	TIRUVANNAMALAI	0.895232	0.456488	-0.732220	Rice	15163	60243	...	84.123502	100.019805	552.942203	615.58
82	Kharif	2020	TAMIL NADU	TUTICORIN	0.895232	0.456488	-0.732220	Rice	327	722	...	84.123502	100.019805	552.942203	615.58
83	Kharif	2020	TAMIL NADU	VELLORE	0.895232	0.456488	-0.732220	Rice	12552	53120	...	84.123502	100.019805	552.942203	615.58
84	Kharif	2020	TAMIL NADU	VILLUPURAM	0.895232	0.456488	-0.732220	Rice	10239	47038	...	84.123502	100.019805	552.942203	615.58
85	Kharif	2020	TAMIL NADU	VIRUDHUNAGAR	0.895232	0.456488	-0.732220	Rice	1	4	...	84.123502	100.019805	552.942203	615.58

86 rows × 22 columns

rspm	spm	Temperature - (Celsius)	State_Name_cat	District_Name_cat	Season_cat	Crop_cat	Pred
52.942203	615.585962	24.644282	0	0	1	0	5.787728
52.942203	615.585962	24.644282	0	1	1	0	5.576660
52.942203	615.585962	24.644282	0	2	1	0	5.134041
52.942203	615.585962	24.644282	0	3	1	0	5.388639
52.942203	615.585962	24.644282	0	4	1	0	6.227887
...	...	...	...	...	...	...	...
52.942203	615.585962	24.644282	0	26	0	0	4.477190
52.942203	615.585962	24.644282	0	27	0	0	3.922375
52.942203	615.585962	24.644282	0	28	0	0	4.390096
52.942203	615.585962	24.644282	0	29	0	0	4.601765
52.942203	615.585962	24.644282	0	30	0	0	4.740211

District Wise Crop yield production is done successfully with respect to factors like season, year, district, production, area, rainfall, production, suspended particulate matter, NO2, S02 present in the soil and lastly temperature. The output is as follows:

```
In [49]: predspredicted[['Season', 'Crop_Year', 'State_Name',
                         'District_Name', 'Max', 'Mean','Min', 'Crop', 'Area','Production', 'Rainfall', 'so2', 'no2', 'rspm', 'spm',
                         'Temperature - (Celsius)', 'Pred ']]

         pspred.copy()

In [60]: fig = px.treemap(p, path=['District_Name','Crop_Year'], values='Pred ')
         fig.show()
```

If we hover through it we would be able to see the crop yield of each district.

End to end product prototype would help the farmers to actually understand and analyze the problems themselves and come up with a solution to improve their yield.

6. Discussion and Conclusion

6.1 Discussion

In this project, we shall also try to investigate the possibility of Crop Rotation in particular piece of farming land by evaluating different parameters like Production, Area of crop field, Temperature, Rainfall, Season, Crop Type and Air Quality Index of historic data.

- Thus we also inferred from our exercise that data transformation is a very important aspect of creating a model which is usually neglected and more importance is given to the hyper parameter tuning of the model.
- This will help the policymakers of the state to determine the budget. If the production of a crop observes a declining trend, then they can plan to implement the schemes at an early stage. This in return will save the state from shortage of a product.

7. Future Plan

- Working on building a Power Bi dashboard for real time analysis of crops. We will a try to build the front end API and link it to the backend Machine Learning model which would predict the particular crop just by entering the temperature of that area and the particular soil requirements. This would be more convenient for the farmers to be very particular about their decision.
- Investigating the idea of crop rotation on a piece of land to be more precise. This could be done by using higher Machine Learning algorithms like CAT Booster, XG-Booster.

References

[1] Thomasvan Klompenburga, Ayalew Kassahuna, Cagatay Cata “Crop yield prediction using machine learning: A systematic literature review”, Computers and Electronics in Agriculture, Elsevier, Volume 177, October 2020, 105709.

[2] Aruvansh Nigam; Saksham Garg; Archit Agrawal; Parul Agrawal, “Crop Yield Prediction Using Machine Learning Algorithms”, Fifth International Conference on Image Information Processing (ICIIP), 15-17 Nov.2019, DOI: 10.1109/ICIIP47207.2019.8985951

[3] Kevin Tom Thomas1, Varsha S2, Merin Mary Saji3, Lisha Varghese4, Er. Jinu Thomas5, “Crop Prediction Using Machine Learning” International Journal of Future Generation Communication and Networking, Vol.13, No.3, (2020), pp.1896–1901

[4] Nischitha K, Dhanush Vishwakarma, Ashwini, Mahendra N, Manjuraju M. R, “Crop Prediction using Machine Learning Approaches” International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol.9 Issue 08, August-2020

[5] Prof. D. S. Zingade, Omkar Buchade, Nilesh Mehta, Shubham Ghodekar, Chandan Mehta “Crop Prediction System using Machine Learning”

[6] Pavan Patil, Virendra Panpatil, Prof. Shrikant Kokate “Crop Prediction System using Machine Learning Algorithms”.

[7] https: //scholar. google. com/scholar?as_q=International+Journal+of+Engineering+Sciences+Research+Technology+Predicting+Yield+of+the+Crop+Using+Machine+Learning+Algorithm&as_occt=title&hl=en&as_sdt=0%2C31

[8] https: //scholar. google. com/scholar?as_q=Applications+of+machine+learning+techniques+in+agricultural+crop+production%3A+a+review+paper&as_occt=title&hl=en&as_sdt=0%2C31

[9] https: //chrisalbon. com/machine_learning/trees_and_forests/decision_tree_regression/

[10] https: //towardsdatascience. com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3
[11] https: //www.w3schools. com/python/python_ml_getting_started. asp