Crop Production Prediction using Machine Learning

Pavitra Kadiyala, Vallabhaneni Sri Harsha Sai
School of Computer Science & Engineering, VIT, Vellore, India

***Abstract*: Agriculture is a very important aspect of one's life. One essential thing for a living being is food. Crop production plays a major role in the amount of food available. Crop production prediction is very important as this will give the farmers and other human beings a rough idea of the prices of food as well as the amount of available food. Machine Learning and Artificial Intelligence help us predicting a few aspects which are difficult for humans to predict. With the help of this domain, we can predict the amount of production of crops in a particular area given the amount of rainfall and the area of land present. This research aims to predict the amount of crop production based on multiple factors such as rainfall, area, location, and type of crop.**

***Index Terms*: Crop Production, Machine Learning, Regres- sors.**

I. INTRODUCTION

Agriculture is the primary source of income for most of the Indians. India has a large production of a variety of crops.

The agricultural industry is growing rapidly and hence the estimation of production is very much important in the field of agriculture. A farmer can decide the pricing based on the production one expects. In India, the amount of rainfall is the main source of water for crop harvesting. Production is directly proportional to the amount of rainfall, though it varies with the area calculated.

The estimation of production with the rainfall is one big challenge when taken over a large area. Different factors determine the production sum of the crop harvested. There are many conventional techniques used to estimate them but not very accurate or reliable. Machine learning is one field that helps us predict things. The data fed to the model can give an accurate result of an estimate of the crop harvested. Hence we are going to train our model on a few parameters to help us predict the production of the crop.

1.1 Main Factors:

- *Amount of Rainfall:* The main essential for crop harvesting is water and in India maximum farmland depends on rainwater. A good amount of rainfall can help in a decent yield of a crop. Hence the amount of rainfall is considered one of the most important factors.
- *Area of land:* The area of the field also plays an important factor. If the amount of rainfall along with the area of the field is taken into consideration, we get the best prediction about crop production. A lesser area with a huge amount of rainfall cannot be favorable for agriculture. The area to rainfall ratio should be of equilibrium values for best production.
- *Type of Crop:* The type of crop needed would also play a crucial role in determining the amount of production. A few crops need more water than the others.

1.2 Alternate Work in this field and their drawbacks.

- *Censuses:* Agricultural censuses, is the most common method of knowing the production, this involves a huge man power, and takes a lot of time to estimate the final value. This method is usually done at a gap of 4-5 years.
- *Local governance data:* The local bodies report the data collected from the farmers to the authorities and the estimate is calculated. Even this involves high manpower and takes a lot of time to come to an estimate.
- *Satellite sensing:* The aerial photographs are taken and the production per area is calculated. This is a very expensive process and requires high-quality image cap- turing.

What our system proposes: Many of the existing features are after the crops are produced. Our model predicts it beforehand. We train the model on existing real time data and predict the amount of crop production. The output is a continuous value and hence we use regressors. The farmer can get the amount of production of crops and can hence get the crop yield. The crop yield is dependent on the amount of production and the area. Hence the farmers can get a rough estimate of their income beforehand. This approach will be faster and has higher accuracy.

II. PROPOSED METHOD

The basic data flow of our research is as follows:

1. Load the dataset
2. Perform Preprocessing(Standardize, Remove unnecessary columns, Dealing with the null values, Character Encoding for Categorical Data)
3. Split the dataset into test and train. Test being the production column and train being the rest of the columns in the dataset.
4. Choose a regressor and train the data on the model

5. Check the R2 Score of the regressor
6. Repeat step 5 and 6 with the rest of the regressor models
7. Choose the model with the best R2 Score.

We have used several Regressors, Supervised Machine Learning Algorithms, for the prediction of crop production. We used an existing dataset. The dataset consists of the region, area, crop type, Rainfall, and Production for the given details. With this, we can predict the crop production details for the ones that aren't present in the dataset as well.

We give the area, rainfall and type of crop as input and get the amount of production as an output. We calculate the R2 score for comparing the models. The dataset we used is referred from 2 datasets. The dataset consists of 74975 entries. The snippet of the dataset:

	State	Year	Season	Crop	Area	Production	Rainfall
0	Andaman and Nicobar Islands	2000	Kharif	Arecanut	1254.0	2000.0	2763.2
1	Andaman and Nicobar Islands	2000	Kharif	Other Kharif pulses	2.0	1.0	2763.2
2	Andaman and Nicobar Islands	2000	Kharif	Rice	102.0	321.0	2763.2
3	Andaman and Nicobar Islands	2000	Whole Year	Banana	176.0	641.0	2763.2
4	Andaman and Nicobar Islands	2000	Whole Year	Cashewnut	720.0	165.0	2763.2

Fig. 1. Snippet of dataset

We did dummy encoding for the categorical data so that no data type error is there. Snippet of the encoding:

```
In [27]: data1 = df.drop(["Year"],axis=1)

In [28]: data_dum = pd.get_dummies(data1)
         data_dum[:5] ##Dummy Encoding
```

Out[28]:

	Area	Production	Rainfall	State_Andaman and Nicobar Islands	State_Arunachal Pradesh	State_Bihar	State_Chhattisgarh	State_Himachal Pradesh	State_Jammu and Kashmir	State_Jharkhand	...
0	1254.0	2000.0	2763.2	1	0	0	0	0	0	0	...
1	2.0	1.0	2763.2	1	0	0	0	0	0	0	...
2	102.0	321.0	2763.2	1	0	0	0	0	0	0	...
3	176.0	641.0	2763.2	1	0	0	0	0	0	0	...
4	720.0	165.0	2763.2	1	0	0	0	0	0	0	...

5 rows × 107 columns

Fig. 2. Snippet of the encoding

2.1 Algorithms

The dataset we used is labeled and hence it is a supervised machine learning method. We trained our model on 7 regressors which are:

1) Random Forest Regressor
2) SGD Regressor
3) LassoLars
4) Bayesian Ridge
5) Passive Aggressive Regressor
6) TheilSen Regressor
7) Linear Regressionode

```
In [32]: from sklearn.linear_model import LassoLars
         model4 = LassoLars()
         model4.fit(x_train,y_train)
         preds4 = model4.predict(x_test)
         r4 = r2_score(y_test,preds4)
         print("R2score when we predict using LassoLars is ",r4)

         R2score when we predict using LassoLars is  0.35571392656117484
```

Fig. 3. Snippet of the code

```
In [37]: print("R2score when we predict using Bayesian  is ",r3)
         print("R2score when we predict using SGD is ",r1)
         print("R2score when we predict using LassoLars is ",r4)
         print("R2score when we predict using PAR is ",r6)
         print("R2score when we predict using TheilSenRegressor is ",r7)
         print("R2score when we predict using LinearRegression is ",r8)
         print("R2score when we predict using Randomn forest is ",r)

         R2score when we predict using Bayesian  is  0.02695906625479294
         R2score when we predict using SGD is  -1.3696811697044418e+22
         R2score when we predict using LassoLars is  0.35571392656117484
         R2score when we predict using PAR is  -0.002397501995262008
         R2score when we predict using TheilSenRegressor is  -0.0008264560215194283
         R2score when we predict using LinearRegression is  0.3557095478780803
         R2score when we predict using Randomn forest is  0.8271022693945856
```

Fig. 4. Output

The figure 3 depicts the code of snippet of one regressor model. The above figure (Fig: 4) depicts the R2 Scores of all the algorithms. We have obtained the highest R2 Score in the Random Forest Regressor. Higher the R2 score, higher possibility of high varience. The next highest R2 score is of LassoLars.

A few points about Random Forest Regressor:

1. Random Forest can be useful for discrete as well as continuous values. In this paper we use it for continuous data.
2. A random-forest regressor is a Meta assessor that fits various grouping choice trees on different sub-examples of the dataset and utilizations averaging to improve the model's accuracy.
3. Random forest is a bagging technique, the trees run in parallel, and there is no interaction between them.

2.2 Large Scale Application

- The model can be further trained for better accuracy and can be utilized by many states. The dataset consists of 12 state's data. The model can be trained on other state's data as well and predict crop production.
- This model can be deployed on a website for large scale purposes. The farmers can then access on their phonesor nearby cyber cafes with ease without any extra money.
- This large scale application can be useful for the citizens as well. The application would predict the crop production and hence the citizens can be aware of the production and can put a rough estimate on the food rates.

III. CONCLUSION

Hence we see that the field of Machine Learning can be useful in the agriculture domain and can help us predict the amount of production of a crop. This can be developed on a higher scale for everyone to use.

REFERENCES

[1] Dataset: District-wise, season-wise crop production statistics from 1997 and District-wise, rainfall statistics.
[2] Priya, P., Muthaiah, U., Balamurugan, M. (n.d.). Predicting yield of the crop using machine learning algorithm. International Journal of Engineering Sceinces Research Technology, 7(4),1-7
[3] Dr. V. Latha Jothi, Neelambigai A, Nithish Sabari S, Santhosh K, 2020, Crop Yield Prediction using KNN Model, International Journal of Engineering research Technology (IJERT).
[4] RTICCT – 2020 (Volume 8 – Issue 12), N. Gandhi, L. J. Armstrong, O. Petkar and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," 2016 13th International Joint Conference on Computer Science and Software En- gineering (JCSSE), Khon Kaen, 2016, pp. 1-5, doi: 10.1109/JCSSE.2016.7748856.
[5] R. Kumar, M. P. Singh, P. Kumar and J. P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique," 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai, 2015, pp. 138-145, doi: 10.1109/ICSTM.2015.7225403.
[6] 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,
[7] Champaneri, Mayank Chachpara, Darpan Chandvidkar, Chaitanya Rathod, Mansing. (2020). Crop yield prediction using machine learning. Inter- national Journal of Science and Research (IJSR). 9. 2.