Crop yield production using machine learning

Chand Kumar Adhaduk
Computer Science & Engineering
PES University
Bangalore,India
adhadukchand@gmail.com

Mahendra N
Computer Science & Engineering
PES University
Bangalore,India
mahendranerella2002@gmail.com

Bhavana Chowdary J
Computer Science & Engineering
PES University
Bangalore,India
bhavanachowdaryjetti420024@gmail.com

***Abstract*— In India agriculture is, no doubt, the major and least paid occupation in the country. Not only that it is the backbone of India. With the help of Machine learning , we can help the farmers by predicting the yield of the crop so that they can plan in advance, so that they can be prepared beforehand. With the help of the model, they can decide which crop to grow so that they can get the required yield. By using attributes like State Name ,Season ,Crop ,Area we will be able to predict the yield of any desired crop.**

Keywords—agriculture, machine learning, yield of the crop

INTRODUCTION

In India, agriculture can be dated back to Indus valley civilization, with its joint sectors, is one of the largest livelihood in India, especially in rural areas of India. Also India ranks second when it comes to agriculture. Total GDP(Gross Domestic Product) of forestry and fishery combined is 15.4 % whereas agriculture alone contributes 31% of GDP which is very large compared to other occupations and also the workforce is 59%. In terms of net cropped area, India comes first and then US and China comes after India. India is an agriculturally developed country.

Predicting the yield of any desired crop using machine learning will surely help. Farmers can plan beforehand and decide which crop to grow. There are a large number of machine learning models which predict the crop yield using different input attributes, some use attributes like temperature and rainfall, while other use static attributes like season, area, state. To make it simple we will be using attributes like crop, area , state, Season to predict the crop yield.

REVIEW OF LITERATURE

In [1] authors predict the yield of almost any kind of desired crops that are grown in India. Their data contains attributes like area, district, state, season and year that are used to predict the crop yield Methodology used by them is Stacked Regression . Its a ensembling technique but a little different ,basically an enhancement of taking mean. In this, author made use of use of meta model and even made use of the a teachnique called out-of-the-fold predictions for the other models which is used to train their primary model wwhich is alo the meta model. Performance metric used by them is Root mean square error(RMSE). Initially RMSE for Enet model was nearly 4%, for Lasso the error close to 2%, Kernel Ridge got nearly 1% and finally after combining all of them gave error less than 1%.

In [2] authors tried to focus on predicting the yield of any desired crop by implementing different types of Machine Learning models. Their dataset contains attributes Temperature and Rainfall .Methodology implemented by them are ensemble learning techniques like KNN Classifier, Logistic Regression, Random Forest Classifier, XGBoost , Linear Regression and Artificial Neural Networks. Performance metric used by them is MEAN ABSOLUTE ERROR .With 22.17 for Simple RNN , with 34.14 for LSTM.

In [3] authors predicted the yield of any desired crop using the ensemble learning technique , in particular Random Forest algorithm based on existing training data set. Data from TamilNadu was used as a training set for construction of the models. Number of attributes were 7 namely Climate Parameters Like Wind Rainfall Humidity Temperature, Soil value of ph, Soil Type, The cultivation expenses ,and Manufacturing Support Vector machines were also used for this project. The methodology used here is CSM.This is helpful as it tracks the seasonal yield of crops and tries to predict based on that. At the end the Random Forest Algorithm is used as it shows good accuracy. Accuracy of the model built was 75%.

In [4] authors predict yield of the crop grown and rate of success for the crops as per values given by the farmer. The data that they came up was for research purpose only and is extracted from Indian government official website, which had records from the year 1997 to 2014 .Maharashtra state was used as a study area. Number of rows were 12000 Columns were DISTRICT,SEASON,CROP,YEAR,PRODUCTION(tons per hectare in lakh) Model used is a 3 Layer ANN having simple linear regression technique with forward propagation and backward propagation. Methodology that were adapted were both encoding and binning for their analysis. Dataset was divided into training and testing set using 80 is to 20 allotment. Model is first trained using basic simple linear Regression technique with artificial Neural Network(ANN). The Learning rate for each of the layer of network is kept constant through out the process i.e. 0.001 reduces MSE by using Adam optimizer, RELU activation function was used for each neuron and gradient descent was used. 82% accuracy was achieved.

In [5] authors predict yield of any desired crop in India using machine learning algorithms. The Dataset used for experiment was collected from government website .Columns of the dataset were , rainfall,production, area under irrigation crop names ,area seasons, and yield from the year the 1950 to 2018. Models that were implemented were machine learning algorithms like from very basic Linear Regression, Decision Tree ,Lasso regression to prevent overfitting , and Ridge Regression to reduce overfitting. The prediction is made for five major crops which are Tobacco, Bajra, Wheat, Jowar,Rice, and Maize. Mean absolute error(MAE),root mean square error(RMSE) were used to have a look at the performance of the model.The prediction is finally obtained using random forest ensemble technique and decision tree . Accuracy that they got is as follows, Linear Regression:89.38% ,Lasso regression:86.33% ,Decision Tree:98.62% , Ridge Regression:89.53%. The Decision tree performs the best with respect to other machine learning techniques.

DATA ACQUISITION3

The selected data has a huge load of information on crop production in India ranging from several years. The data was taken from https://data.world/thatzprem/agriculture-india .
Dataset has 246091 entries and 7 columns in which four are categorical. Based on the information the ultimate goal is to predict the tield of any desired crop production using machine learning techniques.The following attributes must be given to the Dataset:State_Name,District_Name,Crop_Year,Season,Crop ,Area,Production.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 246091 entries, 0 to 246090
Data columns (total 7 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   State_Name     246091 non-null  object
 1   District_Name  246091 non-null  object
 2   Crop_Year      246091 non-null  int64
 3   Season         246091 non-null  object
 4   Crop           246091 non-null  object
 5   Area           246091 non-null  float64
 6   Production     242361 non-null  float64
dtypes: float64(2), int64(1), object(4)
memory usage: 13.1+ MB
```

As we are not working with Crop_Year and District_Name we have dropped it from the dataset.

METHODOLOGY

```
df.isnull().sum()
```

```
State_Name          0
District_Name       0
Crop_Year           0
Season              0
Crop                0
Area                0
Production       3730
dtype: int64
```

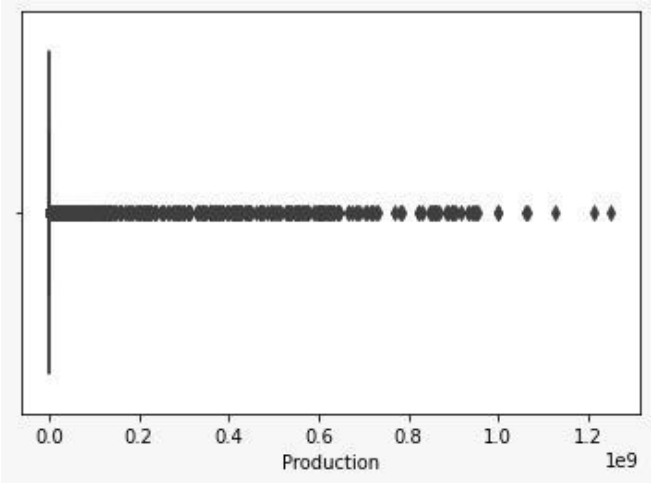
Null values are replaced with Nan values in the data frame and the Nan values are dropped when needed.

```
df.dropna(axis=0,inplace=True)
df.isnull().sum()
```

```
State_Name    0
Season        0
Crop          0
Area          0
Production    0
dtype: int64
```

```
df.shape
```

```
(238838, 5)
```



It's clear from the graph that there are many outliers but removing them wouldn't be a good approach because we would be losing a lot of data. Therefore we decided to remove only the most extreme values.

We also had nominal variables so we had to use one hot encoding which increased the number of attributes. To prevent dummy variable trap we had to remove one of the column. States, Crop and season were the attributes which were one hot encoded.

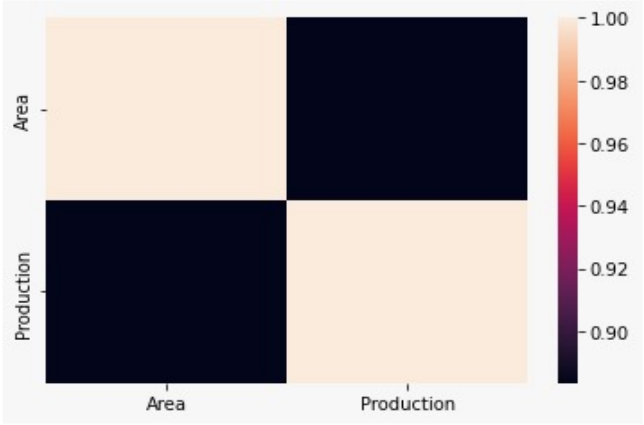
Once the encoding was performed , normalization was performed to scale down the values. So no variable value dominates the other.

As we were dealing with regression problem statement. The ML models which we were gonna use narrowed down. We started from simple models like Simple Linear regression and went up till XGBoost. The metrics we used were r2(coefficient of determination) and rmse(root mean square error).

Linear regression showed bad accuracy , so we tried to refine the linear regression using lasso regression to prevent overfitting and ridge regression. But the accuracy was sufficient. So we decided to use other Machine Learning models like Knn regressor, decision tree regressor, random forest regressor, svm regressor. We got a decent accuracy and error and to refine the model further we applied hyperparameter tuning which further increased the performace. Then we applied Gradient Boosting and XGBoost which are state of the art models. We got the best accuracy and lowest error using these models.

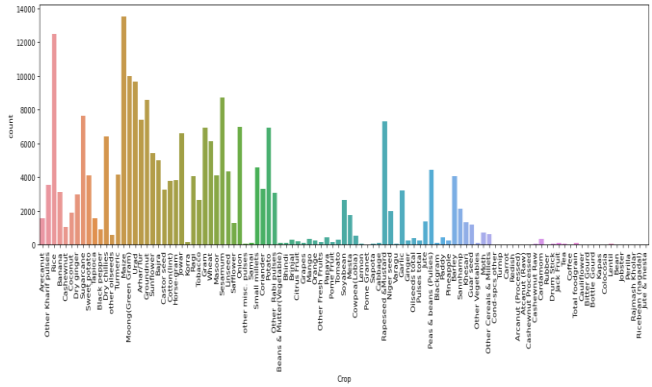
FEW OBSERVATIONS AND INFERENCES:

Area and Prediction columns of the dataset are plotted to check for correlation as shown below:



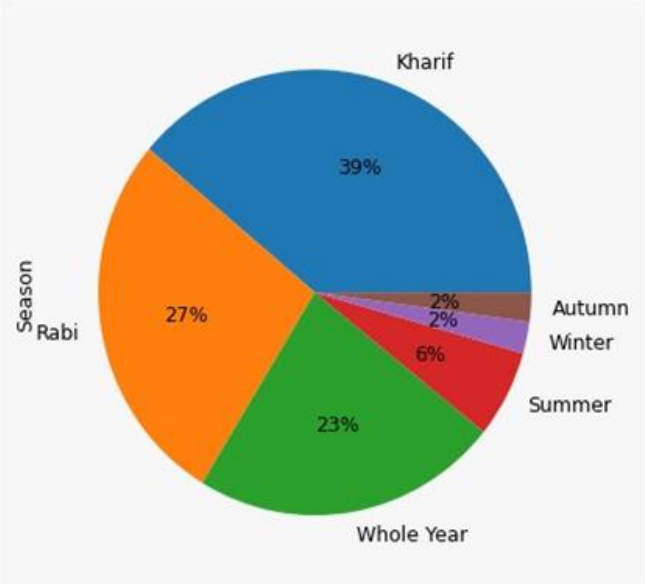
Inference: Area and Production are highly correlated.

A bar plot showing the count of different crops is plotted as shown below:



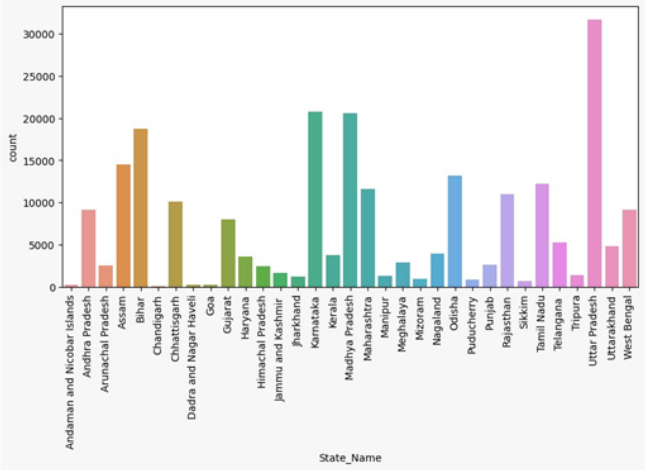
Inference: Top 5 crops grown in India are Maize,Rice,Moong,Urad,Sesamum

A pie chart showing the distribution of crop growth in different seasons is plotted as shown below:



Inference:In most parts of India Kharif and Rabi are the season for cultivation whereas Autumn and Winter are least and In few places crops are grown whole year.

A bar plot showing distribution of crop growth in different states is plotted as shown below:



Inference: Uttar Pradesh, Karnataka, Madhya Pradesh are the top agricultural states in India due to high-yielding genetically modified varieties of seed, also the greater availability of fertilizers and increased use of the facility of irrigation.

Accuracy and RMSE of ML models implemented:

Model	R2	RMSE
Linear Regression	0.218	0.842
Lasso Regression	0.22	0.8422
Ridge Regression	0.225	0.8421
KNN Regressor	0.9025	0.297
SVM Regressor	0.927	0.256
Decision Tree Regressor	0.9231	0.263
Random Forest Regressor	0.9281	0.255

Gradient Boosting	0.9314	0.2493
XGBoost	0.9374	0.2381

Inference: XGBoost models gives the best accuracy(R2) of 93.74%.

ACKNOWLEDGMENT

We would like to acknowledge our Data Analytics Course Professor Dr. Gowri Srinivasa,assistant professors and teaching assistants for providing constant guidance during each phase of our project.

REFERENCES

[1] Jabber B ,Sai Venkat Pinapa ,Sai Nishant Potnuru, Lakshmi Avinash Bollu, "Crop Yield Prediction based on Indian Agriculture using Machine Learning",, India. Jun 5-7, 2020 In:2020 International Conference for Emerging Technology Belgaum.

[2] Agrawal Archit , Nigam Aruvansh, Garg Saksham ,"Crop Yield Prediction Using Machine Learning Algorithms", In:2019 Fifth International Conference on Image Information Processing

[3] Namgiri Suresh,N.V.K.Ramesh,Syed Inthiyaz,P. Poorna Priya,Kurra Nagasowmika,Kota.V.N.Harish Kumar,Mashkoor Shaik, "Crop Yield Prediction Using Random Forest Algorithm",In:2021 7th International Conference on Advanced Computing & Communication Systems (ICACCS).

[4] S. Patil Preeti, S. Kale Shivani, "A Machine Learning Approach to Predict Crop Yield and Success Rate", . Dec 18-20, 2019. In:2019 IEEE Pune Section International Conference MIT World Peace University, Pune, India

[5] Mathur Pratistha, Kavita ,"Crop Yield Estimation in India Using Machine Learning", Oct 30-31, 2020 In:2020 IEEE 5th International Conference on Computing Communication and Automation Galgotias University, Greater Noida, UP, India..