An Efficient Crop Yield Prediction System Using Machine Learning Algorithm

Saiteja Kunchakuri+, SaiShivani Pallerla, Sathya Kande, Nageswara Rao Sirisala

Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India
E-mail: saiteja.kunchakuri@gmail.com+

Keywords: Crop Yield, Agriculture Management, Harvest Production, Yield Prediction, Machine Learning, KNN Regression.

Abstract

Accurate yield estimation is essential in agriculture. Crop yield prediction is required, mainly in countries like India, which rely on agriculture as their prime source of economy. Many developments are required in the agriculture field to increase changes in our Indian economy. One of the major advancements is to implement machine learning in agriculture. Crop yield estimation helps farmers to decide about the crop in the initial stages of agriculture. Machine learning based techniques help for accurate crop yield prediction. The rapid progress in remote sensing technologies and machine learning techniques will give cost-efficient approaches for better crop yield prediction. We used a dataset from a government site for our study in this paper. In this work, using the KNN algorithm an accurate Crop Yield Prediction System (CYPS) is implemented. While for a farmer, predicting yield needs to be more accurate based on various features which could affect yield production and quality. Three factors that affect yield production include season, crop type, and area of production, so to predict the yield production, we are using certain fields like region, year, crop, season, area in our work. Accurate information about the history of crop yield is a crucial thing in creating decisions associated with agricultural risk management.

1 Introduction

Main service of our proposal is to act as a guide/mediator and give virtual support to the farmer and help him/her get aware of investments related to crop and production based on various features. Idea of our proposal is to predict the crop yield based on the factors provided using Machine Learning. This prediction is useful for farmers to get idea about the crop.

Crop yield estimation is a significant farming issue. Farmers require knowledge related to the crop yield before sowing seeds in their farms. A model to predict like this is quite helpful for a farmer who could utilize the information provided daily also. The crop yield mainly relies upon climate conditions, pests, and planning of harvest methods. Accurate knowledge about the history of crop yield is a significant element for making decisions associated with agricultural risk management. Basically, Machine learning is an application of Artificial Intelligence that provides ability to the system to gain knowledge and experience to run automatically, [1].

Machine learning has many branches like classification, clustering, regression etc. Based on our dataset and output values we can say, output is a continuous value so here, we use Supervised Learning [2] and Regression to build the model [3]. The statistical approach in which features are taken as input and target feature is given as output is known as Regression. Coming to our proposal it is clearly visible that target feature is yield and input features are type, season, location etc. Here, one of the best algorithms used is KNN regression. KNN regression gives accurate outputs compared to other models. Fuzzy logic is also one of the techniques used for predictions, [4,5].

The paper is organized as follows. In Section 2, the details to related study is described. In Section 3, Proposed model is explained. In section 4, Results are analyzed. Section 5 concludes this paper followed by future scope.

2 Related Work

According to a paper proposed by Mayank Champaneri [6], a classifier algorithm is used to develop a model to predict yield for any crop field. They used the Random Forest algorithm on the dataset which was collected from various government sites for prediction to achieve great accuracy. Random forest takes more

time to build decision trees and it needs effective resources to build the model.

In a recent paper proposed by Shanmugam Shoba [7], the study of crop yield regression is based on the Multiple Linear Regression (MLR) algorithm. They used regional data from the Methodological department of the Mysore district. They stated that their model has an accuracy of 91% for seasonal crops and 74% for yearly crops. But the main disadvantage is with the MLR algorithm because Multiple Regressions are based on assumptions that they have a linear relationship between dependent variables and independent variables. Nature of data plays a key role while using MLR algorithm, rigorously observed data is best suitable for MLR algorithm.

The paper proposed by David Lobell [8] suggests that the use of a deep learning framework along with remote sensing data gives accurate predictions. They implemented CNN and LSTM on histograms to predict crop yield.

According to Shilpa Mandal [9], where crop yield prediction is done using deep learning. Deep learning gives the accurate yield. However, it is very complicated and requires more libraries compared to our model and also require more dataset values in deep learning. It is also expensive to train as it contains complex data models. There is a need of large and expensive software systems and also lot of machines. No standard theory is present for selecting proper and preferable deep learning tools. This makes difficult for less knowledge people.

In a paper, Kanwal Preet Singh Atwal [10] designed a model using the Apriori algorithm. They analyzed the effect of temperature and rainfall, especially on paddy yield using the WEKA tool. The research was done on the dataset taken from government sites, performed data preprocessing, data discretization, and used for further study.

As stated by Saeed Khaki [11] in a paper, using a deep neural network gives accurate predictions. They used data from Syngenta Crop Challenge which is publicly available for study in yield prediction. Although the model predicts precisely the major disadvantage is the black box property and they stated that the DNN models become less explainable sometimes.

In a paper proposed by M Suganya [12], they compared various supervised learning techniques. They used remote sensing data from satellite imagery for their study. They compared different classifiers and regression models and concluded Logistic Regression is best for their data.

Palanivel Kodimalar [13] in a paper suggested applying various machine learning algorithms for predictions and recommendations using big data techniques. They recommended that big data has a great impact on agriculture. They used big data environment for preprocessing and prediction.

Devdatta A. Bondre [14] analyzed the features required for agriculture in his paper to predict the yield production. The research was conducted on selected crops and data collected from different sources. He mainly focused on soil classification and fertilizer recommendations based on the soil nutrients and past data of the crop.

In a paper by K Porchilambi [15], they explained in detail how to use different machine learning algorithms for crop yield prediction. They focused on Naïve Bayesian, Support vector machine, Neural networks, Decision tree, K Nearest Neighbor for their study using dataset collected from Tamil Nadu government website.

3 Crop Yield Prediction System (CYPS)

Machine Learning has set foot in the medical field [16], it has a lot to do in the agriculture sector. Our model can help to predict the crop yield to cut down the complications faced by the farmers. Currently, Remote sensing (RS) models are comprehensively used in building decision making tools for agricultural systems to enhance the estimations of crop yield. More coverage is possible with less environmental impact. Remote sensing-based techniques need the processing of an immense number of remotely sensed information from various sources. Hence, significant consideration is currently being given to machine learning (ML) techniques. This is because of the capacity of Machine learning based models to deal with a huge number of inputs and manage non-linear tasks.

KNN Algorithm is used to build our proposed crop yield prediction model. In this model, we consider a query crop for which yield is to be predicted. K crops that are near to features of query crop are taken from the dataset. Using these k crops, the yield production of these crops is extracted from the dataset. The average of these k yield production is nothing but the yield production of the query crop. The advantage of the proposed model is its high performance and accuracy rate which is flexible.

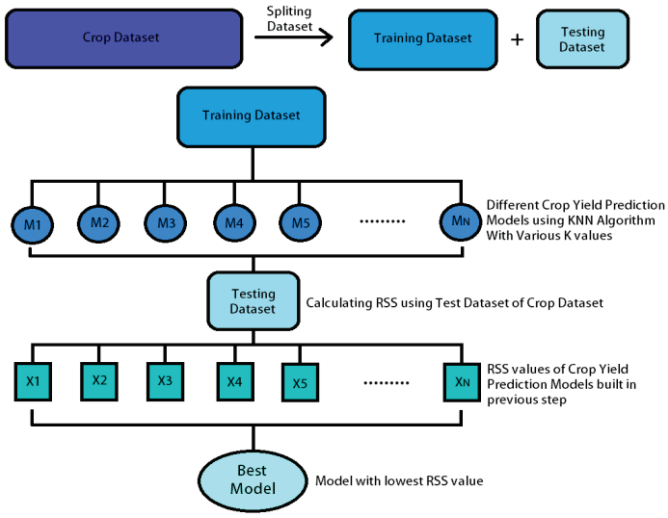

Figure 1: Flow graph of CYPS algorithm.

The working principle of our model is shown in figure 1. The best k value is taken from a range of values. Initially, the dataset is split into training and testing datasets. And using the training dataset various KNN models were built. The RSS value of all these models is calculated using the test dataset. The model with the least RSS value is considered as the best predictive model among all models. By using this algorithm dataset is trained and knowledge is given to the system to predict the crop yield. After gaining knowledge from the system, when a particular set of factors are given as an input, the yield production of the crop in tons comes as output.

CYPS Algorithm:

(1) Initialize two lists Kdist[] and KCrop[] such that,
Kdist= list of first k crop's Euclidian distance from query crop in sorted order=sort(δ1,……δk)
KCrop= list of first k crops in sorted order based on distance=sort(c1,…ck)
(2) Initialize i=k+1
(3) Compute δ=distance between C_i and Cq(query crop)
(4) Go to STEP(5) if δ<Kdist[k] else Go to STEP (7)
(5) Find variable j such that δ>Kdist[j-1] but δ<Kdist[j]
(6) Then remove furthest crop and shift queue:
KCrop[j+1:k]= KCrop [j:k-1]
Kdist[j+1:k]=Kdist[j:k-1]
Set Kdist[j]=δ and KCrop [j]= C_i
(7) Then increment i by 1
(8) Go to STEP (3) if i<=N
(9) Return KCrop. Now this gives the list of k closest crops to query crop.

As part of working process of KNN Algorithm after choosing the K value, the Euclidean distance is calculated between the query and other crop attributes from the training dataset. After this, sorting of those distances is done. So based on this distance array average of K nearest crops yield is considered as the result.

4 Result Analysis

Choosing the dataset required is one of the challenging tasks. Even after choosing data, feature analysis is important and the features that we required should be extracted. So, we considered only the features that influence the yield production.

Attributes of the dataset used in our study are listed below:

- Crop Type
- Season (Kharif, Rabi, Whole Year)
- Location (State, District)
- Area Production (in Hectors)
- Yield production (in Tons)

The target attribute considered for our study is yield production in tons. Yield production is different for the different crops so that crop type plays a vital role in yield prediction. Mainly production of a crop depends on rain so using the season feature yield can be predicted accurately. The feature location influences production because for a particular crop based on geographic location yield production changes due to variations in multiple factors like soil, temperature, humidity, etc. sample data from the dataset can be observed in table 1.

Id	1	2	3	4
State name	Andaman & Nicobar	Bihar	Gujarat	Odisha
District name	N And M Andaman	Purbi	Patan	Bargarh
Crop year	2010	2010	1997	2007
Season	Rabi	Kharif	Kharif	Winter
Crop	Dry chilies	Mesta	Bajra	Sugarcane
Area	254.0	7.0	85900	742.0
Production	402.0	120.0	103000	47581.5

Table 1: Sample data from the dataset.

After getting the dataset, preprocessing should be done so that our prediction won't go wrong. Null values should be checked and

removed using inbuilt functions. Feature outliers should be checked using boxplots. If outliers are found in the dataset, they should not be considered for further study. Using box plots outliers at attribute level can be observed. Exploring data or data analysis is a fundamental step before building a predictive model [17]. Researchers can identify the distinct patterns of the data by studying data. Distribution of production can be observed in the figure 2 over the years.

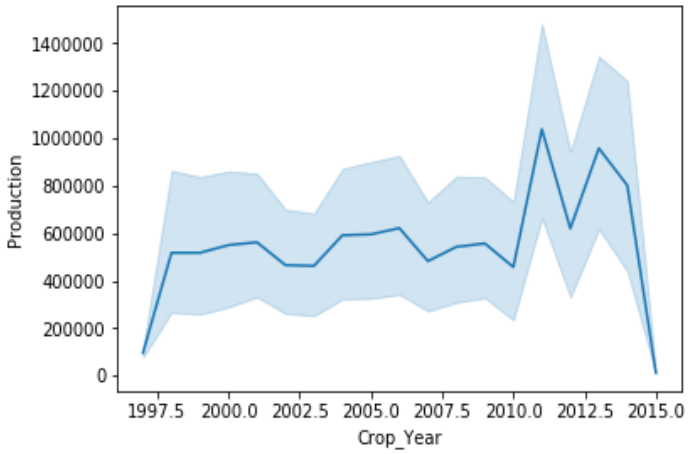

Figure 2: Distribution of crop production over years.

Dataset is split into two parts; one is the training dataset (75%) and the rest is the testing dataset (25%). The training dataset is used to build the model whereas the testing dataset is used to evaluate the model. Euclidian distance is calculated between to integer values. Strings are handled by replacing them with integers so that calculating Euclidian distance will be easy. Vectorization functions in python language help to replace strings with integer values as shown in table 2.

Id	1	2	3	4
State name	0	4	9	22
District name	1	87	88	54
Crop year	2010	2010	1997	2007
Season	3	0	0	4
Crop	11	57	20	7
Area	254.0	7.0	85900.0	742.0
Production	402.0	120.0	103000.0	47581.5

Table 2: Data after Preprocessing.

Euclidian Distance is used to calculate the distance between any two crops (In computing distance it is crucial to normalize features otherwise features with higher order will influence the distance more than the features with lower order). Dataset values should be normalized such that they fall in the range (0,1). Model in our paper is built such that it gives the average of first k nearest neighbors based on Euclidian distance which is calculated using the formula in Equation (1).

$$d(p, q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

p, q = crops for which distance is calculated

n= no of attributes

Different models are built using different k values and the best model among them is selected by comparing their RSS (Residual sum of squares) value. Residual Sum of Squares is metric we used in study to evaluate different models and calculated using Equation (2). Model with the low RSS is better model then the model with high RSS.

$$RSS(\beta) = \sum_{i=1}^{N}(Y_i - \hat{Y})^2 \quad (2)$$

Yi = Actual Output $\hat{Y}$ = Predicted value

N= no of results

Here various models and predictions of the proposed model are analyzed using some test cases. For KNN, we need to find a set of k crops from the training dataset which are closer to the given test case crop by comparing Euclidian distance. Then calculate the average of these k nearest neighbors. There remains a question of choosing the value of k to use in making predictions. For our study using the KNN algorithm the best k value should be chosen based on RSS value in range 1 to 16.

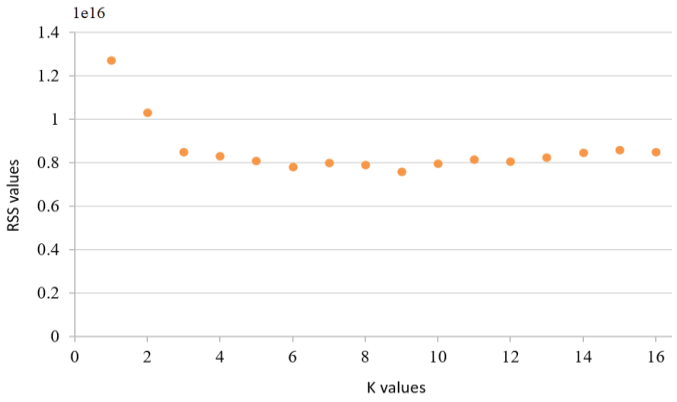

Figure 3: RSS values of CYPS Models with different K values.

From the above graph of RSS values, we can say that it decreases with the increase in the k value. However, after a threshold value, the RSS value fluctuates. Point is to take the k value which gives the lowest RSS. From figure 3 we can say that the RSS value is less at K value 9 among others. This value depends on the dataset taken this is not the same for every dataset.

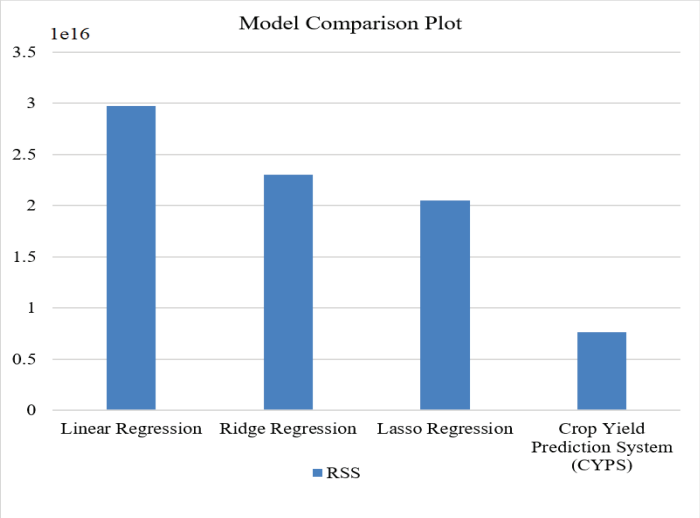

Figure 4: Comparing various models.

Comparison between linear regression, Lasso regression, ridge regression, and KNN is done to find the best predictive model. The RSS value of the KNN algorithm is much smaller than other regressive models as shown in figure 4. So KNN is considered the best model for crop yield prediction.

5 Conclusions

This research confirms that the selected paper uses various attributes, providing the scope of the analysis and the accessibility of information. Crop yield prediction can help farmers to increase productivity and profitability. It assists farmers in planning their agricultural cycle perfectly and precisely. Research show that many systems with numerous features did not generally give the optimal performance for crop yield prediction. Therefore, we used the best result-giving features to build the model which gives the best prediction and increases the performance compared to the already existing models. Although, there are different kinds of algorithms used for crop yield prediction. By mentioning the advantages of this model and also the easiest way to build it, we can also conclude that the KNN algorithm is more accurate compared to other models. We consider that this paper will be helpful for further analysis on the advancement of crop yield prediction systems. In the future, we can determine an efficient algorithm based on their accuracy metrics that will help to select an efficient algorithm to predict crop production.

Fields like humidity, weather conditions, soil nutrients, soil type, soil quantity, and many more factors can also be included in future projects for higher accuracy. There are different algorithms in deep learning where in the future we can use for better prediction of crop yield. By suggesting more profitable crops. In the future, we can also build recommendation systems for a given set of input as factors. To make users more comfortable we can also develop a user interface and can also build some easy applications to use in mobile phones.

References

[1] Saranya C P, Guru Murthy B, Karuppasamy M, Sunmathi M, Shree Sakthi Keerthna S, “A Survey on Crop Yield Prediction Using Machine Learning Algorithms”, ©2020 IJRAR March 2020, Volume 7.

[2] Manoj G S, Prajwal G S, Ashoka U R, Prashant Krishna, Anitha P, “Prediction and Analysis of Crop Yield using Machine Learning Techniques”, ©IJERT, NCAIT – 2020, Volume 8, Issue 15.

[3] P.Priya, U.Muthaiah, M.Balamurugan, “Predicting Yield of The Crop Using Machine Learning Algorithm”, © (IJESRT) International Journal of Engineering Sciences & Research Technology, ISSN: 2277-9655.

[4] Nageswara Rao Sirisala, C.Shoba Bindu. “Uncertain Rule Based Fuzzy Logic QoS Trust Model in MANETs”, International Conference on Advanced Computing and Communications - ADCOM, (IITM PhD forum), (2015), pp.55-60.

[5] Nageswara Rao Sirisala, C.Shoba Bindu. “A Novel QoS Trust Computation in MANETs Using Fuzzy Petri Nets”, International Journal of Intelligent Engineering and Systems, Vol.10, No.2, (2017), pp 116-125.

[6] Mayank Champaneri, Darpan Chachpara, Chaitanya Chandvidkar, Mansing Rathod, “Crop Yield Prediction Using Machine Learning” International Journal of Science and Research in April 2020.

[7] M Shanmugam Shobha, Vijay Hegde S, Yashvanth C V, S Chandra Kiran, “Crop Yield Prediction Using Machine Learning Algorithm” ©2020, International Research Journal of Engineering and Technology, Vol 2, ISO 9001:2008.

[8] David Lobell, Jiaxuan You, Xiaocheng Li, Melvin Low, Stefano Ermon, “Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data”, AAAI Conference on Artificial Intelligence (AAAI-17).

[9] Shilpa Mandal, Rohan Yadav, Satyam Yadav, Neha Gunjal, “Agricultural Crop Yield Prediction using Deep Learning Approach”, ©IRJET, Volume: 06 Issue: 12, Dec 2019.

[10] Kanwalpreet Singh Attwal, Kuljit Kaur, “Effect of Temperature and Rainfall on Paddy Yield Using Data Mining”, ©IEEE**,** 7th International Conference Confluence 2017 On Cloud Computing, Data Science and Engineering.

[11] Saeed Khaki, Lizhi Wang, “Crop Yield Prediction Using Deep Neural Networks”, Frontiers in Plant Science, May 2019, Volume 10, Article 621.

[12] Suganya M, R Dayana, R Revathi, “crop yield prediction using supervised learning techniques”, June 2020, International Journal of Computer Engineering and Technology, 11(2), 2020, pp. 9-20.

[13] Palanivel Kodimalar, Surianarayanan Chellammal, An Approach for Prediction of Crop Yield Using Machine Learning and Big Data Techniques (2019). International Journal of Computer Engineering and Technology 10(3), pp. 110-118, 2019.

[14] Devdatta A. Bondre, Santosh Mahagaonkar, “Prediction of Crop Yield and Fertilizer Recommendation Using Machine Learning Algorithms”, ©IJEAST, 2019 Vol. 4, Issue 5, ISSN No. 2455-2143.

[15] K Porchilambi, Dr P Sumitra, “Machine Learning Algorithms for Crop Yield Prediction: A Survey”, Journal of Emerging Technologies and Innovative Research, Volume 6, Issue 3, March-2019, eISSN: 2349-5162.

[16] M. S. Raja, M. Anurag, C. P. Reddy and Nageswara Rao Sirisala, "Machine Learning Based Heart Disease Prediction System," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-5.

[17] Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, Subhendu Kumar Pani, “Exploratory Data Analysis using Python”, IJITEE, ISSN: 2278-3075, Volume-8, Issue-12, October 2019.