Performance Evaluation for Crop Yield Prediction Using Machine Learning Algorithms

Nayan Ranjan Das
2nd Year, Dept of Computer Science and Engineering
Amity School of Engineering and Technology
Noida, Uttar Pradesh, India
nayan.das@s.amity.edu

Garima Sharma
2nd Year, Dept of Computer Science and Engineering
Amity School of Engineering and Technology
Noida, Uttar Pradesh, India
garima.sharma@s.amity.edu

Dharya Pratap Singh
2nd Year, Dept of Computer Science and Engineering
Amity School of Engineering and Technology
Noida, Uttar Pradesh, India
dharya.singh@s.amity.edu

Guide: Dr. Shilpi Sharma
Faculty, Dept of Computer Science and Engineering
Amity School of Engineering and Technology
Noida, Uttar Pradesh, India
ssharma22@amity.edu

***Abstract*— This paper mainly discusses the fusion of technology with the agricultural sector of India, which has contributed and will contribute extensively to GDP, and is responsible for the majority of employment in the country. We have used Machine Learning (ML) techniques like Curvilinear Regression (CLR) for the quantitative analysis, and K-Nearest Neighbours (KNN) classification for the qualitative analysis of crop yield. Previously, the work in this field was based just on the quantitative analysis of the data to plot the relationship of Simple Linear and Multiple Linear Regression. Artificial Neural Network (ANN) has also been repetitively used to find the results which are more qualitatively oriented. The factors used by various researchers are varying i.e., different from each other. Thus, these proposed models have volatile predictions and varying accuracy rates. The extrapolation is mainly based on the various factors like Soil conditions, Area under Control (AUC), Annual Rainfall (AR), and it also affects the crop prices i.e., the Minimum Support Price (MSP). It might result in the fluctuation of the market. Classification done using KNN branches the available data into respective categories using all the given parameters. We have used the parameters like: Land Area being used for agriculture and the States. We have proposed and implemented a model to carry out accurate and efficient crop yield prediction.**

***Keywords* - Crop Yield Prediction, Machine Learning, Linear Regression, K-Nearest Neighbours, Supervised Learning.**

I. Introduction

Agriculture is an indispensable pillar for the Indian Economy as it contributes approximately 20% [1][2] i.e., ⅕ of the Indian GDP and is responsible for the employment of 152 million [1] which includes 15.9% increase from the previous years, even after the pandemic.

India is the second largest producer of crops [3]. One of the major adversities faced by farmers is the crop yield prediction due to uncertainties in Weather conditions and the rainfall, cost and availability of fertilizers, type and level of nutrition in soil, and several other factors leads to deprecation in the production of crops.

The main aim of this paper is to create a simple and easy illustration for the farmers, so that they are able to determine the crop yield expectation. Indian Agriculture data is enormous and when this data becomes information, it can solve many problems [3].

The dataset undergoes cleaning and pre-processing, and that is how it gets ready for the purpose of implementation of the model and concluding in accurate crop yield results.

Regression Analysis associates all the factors influencing the yield and its production. It mainly considers the quantitative aspects to create a relation and plot the desired graph. The K-Nearest Neighbour methods applied help us to classify the types of crops state-wise.

II. Literature Review

As referenced from research paper [3], it mainly has incorporated a hybrid MLR-ANN model to calculate the yield. As the ML techniques are more data driven, thus, the correctness of the predictions highly dependent on the data quality. The predictive power reduces with the introduction of error and noise. The model measures the optimal minimum error and this results in increase of prediction accuracy over the same data. Another paper [2] has implemented MLR along with a Density based clustering technique to develop a user-friendly platform to maximize the crop yield.

As discussed in research paper [4], ANN is considered as a better method to interpret crop variability. Mostly the models use the linear affiliation to anticipate the crop production using the enormous data available, without including the fuzzy relations. ANN model is capable of dealing with the

qualitative data, covering the complex, non-linear relationships with the parameters, and overcoming the drawbacks of other models.

In this paper [5] authors discuss the effect of climate-change on crop yield by building a numerical model that focuses on main crop growth and development processes. This paper used the Ceres-Maize model, a process-based crop model. The advantage of using this model is that it calculates true response to climate change. The only limit on the approach of this model is that it only represents some of the numerous processes impacting yield. The dataset includes data gathered over 39 years from 198 sites. The statistical model improves as the scale of interest becomes broader.

The paper [6] discusses a machine learning method for anticipating crop yield and its success rate. The paper has focused on Artificial Neural network (ANN) technology for the prediction. They have used pH, nitrogen, oxygen & potassium content as parameters. They have used a multilayer-perceptron model, developed using a neural network. The model proposed uses backpropagation to reduce Mean Squared Error (MSE) by using RELU activation function and gradient descent. As the number of epochs increases, error reduces. The conventional linear regression techniques are less precise than the methods utilising the technique of ANN. The ANN model accuracy can improve by including more layers and parameters. This paper [7] embraced a very analogous tactic where the attributes selected were pH, nitrogen content, depth of percolation, range of temperature in the region and rainfall. They too have utilised computation of MSEs in their illustration to determine the appropriate integer value of hidden layers in their ANN structure.

The paper [8] discusses forecasting crop yield using classification techniques. It also highlights the data mining process. The techniques used are Classification, Association rules, clustering and Regression. Naïve Bayes, J48, Random Forest, Artificial Neural Network, Decision tree & Support Vector Machine are the algorithms used for classification. The prediction is made by passing climate and crop parameters in the classification model.

In this paper [9] authors discuss coupling machine learning and crop modelling for improving crop yield prediction, in particular corn. They have designed five Machine Learning models – linear regression, LASSO, LightGBM, random forest, and XGBoost, and six ensemble models for corn yield prediction. The observations concluded that adding Agricultural Production Systems sIMulator (APSM) with other climate and soil parameters as input parameters, increased the accuracy of prediction between 7 and 20%.

As referenced in this research paper [10], the three independent factors considered are - Annual Rainfall (AR), Area Under Control (AUC), and Food Price Index (FPI) to profess the impact on crop yield using the Linear Regression analysis. These factors are mainly interlinked and in turn affect each other. The gap discussed in this paper is to increase the horizon of the parameters involved like Weather Conditions, Minimum Support Price (MSP), Soil Parameters.

In this paper [11], deep neural networks have been used, to predict yield of corn hybrids using data on genotype of the crop and environment. Deep neural network (DNN) is a part of ANN, which has multi-layers between the input and output. As network goes deeper, more features can be included to get improved accuracy. Deep neural network models are known to estimate most of the function, making its use versatile. They trained two deep neural networks and then used their outputs to predict the yield. The study goal is to compare the three models – DNN, SNN & Lasso. Their output difference was used for the prediction of response. The result concluded that DNN outperformed all other models in every benchmark.

The paper [12] have made use of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for building a model to predict crop yield based on environmental data and management practices. CNNs are powerful image processing, artificial intelligence (AI) that use deep learning to perform both generative and descriptive tasks. CNNs, are powerful AI tool that uses deep learning to process signals, sequences, images and videos. RNN is a subset of ANN where a directed graph is formed using connections between nodes along a

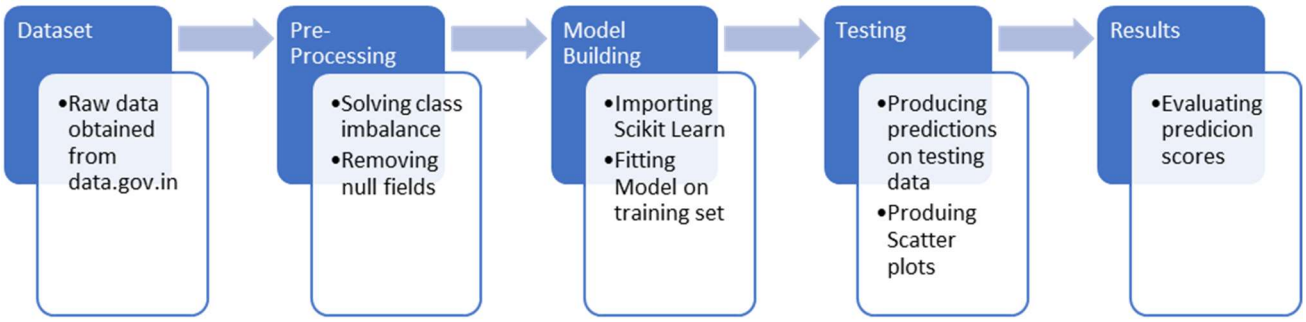

Fig. 1: Depiction of the model-building pipeline

temporal sequence. A Fully Connected (FC) layer was used to combine the features of weather and soil, which was extracted by W-CNN and S-CNN. They also implemented three more models, Random Forest (RF), Deep Fully Connected Neural Network (DFNN) & LASSO. The result demonstrated that hybrid CNN-RN model outperformed all other models. Another insight from this literature that radiation was the most and snow the least sensitive factor in crop yield prediction.

This paper [13] focuses on predicting the crop yields using a Decision Tree model constructed using C4.5 algorithm, implementation of Decision Tree in Weka, a well-known Machine Learning platform. Along with a prediction model, they have developed a web-based platform to host their model. The researchers used data regarding crops grown in Madhya Pradesh. It highlighted a trend where production of each crop is severely impacted by the effects of climate change.

Using IOT as discussed in [14] has broadened the field horizon vision using the multispectral-imaging and remote-sensing. Easy transfer of data is promoted through IOT as the data collected is deployed over GSM network. One of the gaps is the non-utilization of drones to enhance the results

III. Methodology

The methodology for construction of the model and obtaining successful predictions from it was a straight-forward pipeline. It started with obtaining the raw dataset, followed by pre-processing and data cleaning. Once the dataset is suitable for our purpose, the model is implemented upon it to derive conclusions.

The dataset used in our model is obtained from data.gov.in, an Open data initiative of Government of India. The platform is hosted by the National Information Centre (NIC), Ministry of Electronics & Information Technology. As per the description of the data provided by the ministry - the dataset on crop yields is a comprehensive collection of data from the year 1997 to 2015, comprising of production of various crops corresponding to states, districts, seasons and area cultivated. It is useful for understanding the relationship between different aspects and their effect on the production. The data also helps to assess various agricultural provisions provided throughout the span of two decades and across different regions of the country. [15]

The column in concern, i.e., the value to be predicted is under label "Production". It refers to the quantity (in metric tonne) of a particular crop produced in given year in a district. The dataset includes two hundred thousand plus data points in the raw dataset. The rows showing empty values are eliminated to obtain a non-null dataset.

The given dataset is severely imbalanced as out of these 246,091 data points, only 434 are suitable to be used in a regression model. This is because the remaining 245,657 points are accumulated in the vicinity where the crop yield is below 0.019×10^9 tonnes. The data chosen are spread nicely in the vicinity of 0.019×10^9 to 1.2508×10^9. Thus, in order to maintain the comprehensibility of the model and to achieve interpretable inferences from it, we have excluded the data points which had value under the label "Production" less than 0.019×10^9 tonnes. With this, the cleaned and pre-processed dataset can be utilized for training the regression model.

For the machine learning models, we use the popular ML python library – Scikit-learn [16]. The various modules and tools bundled in the library are quite efficient. We have chosen Curvilinear Regression as our regression analysis tool. In Linear Regression, the response vector is linearly dependent on the independent input attributes. Thus, as a result of their linear relationship, we obtain a straight line (governed by equation: $y = A + Bx$, for y as the predicted response, A and B as the coefficients of regression and x as the input variable) as the regression line. Although such approaches are easier to comprehend, these fail to interpret real life data which is not always defined linearly. In the case of Curvilinear Regression, the relationship is defined by a polynomial of some degree greater than 1. For this paper's results we have set the degree of model as 2

Our group divided the processed data into training set and validating set using *train_test_split()* function from the Scikit-learn library. The Regression model is fitted upon the training set, i.e., the input attribute "Area" and the corresponding response "Production". Upon fitting the model, we can obtain the coefficients of regression. Since the degree of the model was selected as 2, the regression line is defined by the equation: $y = A + Bx + Cx^2$. Once the model is prepared and the coefficients have been obtained, the model can be tested on the testing dataset. Using the input vector "Area" from the testing set, predictions can be made for the response label "Production". These predictions are used to plot the regression line on the dispersed plot of the testing dataset.

We use the same processed and undivided dataset for classification analysis. We wish to predict the production the on grounds of state. But the response vector is quantitative. Hence, we encode the column "Production" owing to the range of yield. Each interval is associated to a class, starting from 0.019×10^9 and size of each interval is 0.2×10^9. After encoding, we get 7 classes: class 0 through 6. The number of data points in each class is presented in Table I. Now, the input attribute "State_Name" is containing string values, but the Scikit-learn classifier for KNN classification cannot accept string values as input vectors. Thus, we need to encode the input attribute column as well using the *LabelEncoder* class. Number of rows for each state is presented in Table II.

For K-Nearest Neighbours classification, the K-value - number of neighbours to be assessed for classification must be specified before deployment of the model. To maintain the simplicity of the model we have put the said parameter as 10 instead of using a complex calculus-oriented algorithm to obtain a value for the parameter. Again, the dataset is split and the classifier is fit on the training set. We can obtain a vector of predicted responses using the method *KNeighborsClassifier.predict()* on the model.

IV. Result

The effectuation of the model can be assessed using many criteria. The veracity of the predictions made by the regression model can be assessed visually from the scatter-regression line plot (Fig. 1). We can see how the best-fit line follows the data points closely, giving errors as minimum as possible on the set degree. Another method to obtain a numeric value for accuracy of the said regression model is to use the R^2 score for regression analysis. This metric is available in the *LinearRegression* class of Scikit-learn library. Using this, we get an estimated value of 0.8222080491879048 for R^2. The limiting significance of this metric is 1. Hence, we can assess that the regression model is capable of predicting the response vector "Production" with high accuracy.

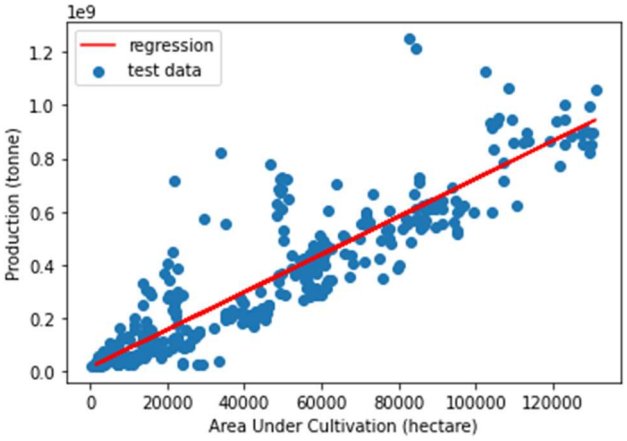

Fig. 2: Test data dispersion plot and anticipated regression line

TABLE I
NUMBER OF ROWS FOR EACH CLASS

Class	Number of data points
0	216
1	78
2	69
3	40
4	25
5	4
6	2

TABLE II
NUMBER OF ROWS FOR EACH STATE

State	Number of data points
0	14
1	83
2	18
3	8
4	238
5	5
6	55
7	13

TABLE III
AN INSTANCE OF CLASSIFICATION REPORT FOR KNN CLASSIFIER

Class	Precision	Recall	F1-score	Support
0	0.93	0.93	0.93	56
1	0.38	0.36	0.37	14
2	0.54	0.35	0.42	20
3	0.41	0.78	0.54	9
4	0.50	1.00	0.67	5
5	0.00	0.00	0.00	3
6	0.00	0.00	0.00	2

The predicted response vector from *KNeighborsClassifier.predict()* is exercised to attain a report of metrics describing the exactitude of the model for each class. The metrics obtained are: precision, recall, f1-score and support. One instance of the report is presented in Table III. The values are amply high for class 0 only, because as per table, class 0 has the maximum amount of data points in the dataset. This imbalanced distribution affects the prediction of the remainder of the classes. This is the reason classes 1 through 6 have very low scores as they do not have enough data points, with class 5 and 6 having only 4 and 2 rows respectively.

V. Conclusion

We can conclude that the machine learning models proposed in the paper have performed fairly, with the Curvilinear Regression model showing remarkable performance in anticipating the crop yield with respect to area under cultivation. The K-Nearest Neighbour classification model works reasonably fine as well for predictions with respect to states, except for the fact that the chosen dataset still had some imbalance, due to which the prediction was biased. Using this set of machine learning models, we can say that although usage of multiple attributes and deploying complex algorithms (Multiple Linear Regression and Artificial Neural Networks) may make the model very close to the actual data, but even simpler approaches like Linear Regression and KNN classification may produce appreciable models which are able to predict with fair success.

VI. Future Work

The quarters which necessitate improvements involve the way the dataset is selected. Instead of omission of data points, we can apply oversampling over the dataset using SMOTE approach. This way we would not lose essential data describing the relationship of crop yield over various attributes. The same can be applied for dataset used for KNN classification to reduce the bias within the classifier. The number of attributes used can also be increased to model the real-life data more accurately in future renditions of this paper. Another area of improvement would be use of an algorithm and Error vs K-value plot to determine the optimum number of neighbours for KNN classifier. This would compose the classification model even more efficient.

References

[1] "Provisional Estimates of Annual National Income, 2020-21 and Quarterly Estimates (Q4) of Gross Domestic Product, 2020-21." *Press Information Bureau*, Ministry of Statistics & Programme Implementation, Government of India, 31 May 2021, www.pib.gov.in/PressReleseDetailm.aspx?PRID=1723153&msclkid=3a4504c1bca411ecad785f1553e44b70.

[2] Ramesh, D., and B. Vishnu Vardhan. "Analysis of crop yield prediction using data mining techniques." *International Journal of research in engineering and technology* 4.1 (2015): 47-473.

[3] Gopal, PS Maya, and R. Bhargavi. "A novel approach for efficient crop yield prediction." *Computers and Electronics in Agriculture* 165 (2019): 104968.

[4] Khairunniza-Bejo, Siti, Samihah Mustaffha, and Wan Ishak Wan Ismail. "Application of artificial neural network in predicting crop yield: A review." *Journal of Food Science and Engineering* 4.1 (2014): 1.

[5] Lobell, David B., and Marshall B. Burke. "On the use of statistical models to predict crop yield responses to climate change." *Agricultural and forest meteorology* 150.11 (2010): 1443-1452.

[6] Kale, Shivani S., and Preeti S. Patil. "A machine learning approach to predict crop yield and success rate." *2019 IEEE Pune Section International Conference (PuneCon)*. IEEE, 2019.

[7] Dahikar, Snehal S., and Sandeep V. Rode. "Agricultural crop yield prediction using artificial neural network approach." *International journal of innovative research in electrical, electronics, instrumentation and control engineering* 2.1 (2014): 683-686.

[8] Sujatha, R., and P. Isakki. "A study on crop yield forecasting using classification techniques." *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*. IEEE, 2016.

[9] Shahhosseini, Mohsen, et al. "Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt." *Scientific reports* 11.1 (2021): 1-15.

[10] Sellam, V., and E. Poovammal. "Prediction of crop yield using regression analysis." *Indian Journal of Science and Technology* 9.38 (2016): 1-5.

[11] Khaki, Saeed, and Lizhi Wang. "Crop yield prediction using deep neural networks." *Frontiers in plant science* 10 (2019): 621.

[12] Khaki, Saeed, Lizhi Wang, and Sotirios V. Archontoulis. "A cnn-rnn framework for crop yield prediction." *Frontiers in Plant Science* 10 (2020): 1750.

[13] Veenadhari, S., Bharat Misra, and C. D. Singh. "Machine learning approach for forecasting crop yield based on climatic parameters." *2014 International Conference on Computer Communication and Informatics*. IEEE, 2014.

[14] Garg, Gauri, et al. "Crop productivity based on IoT." *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*. IEEE, 2017.

[15] "District-Wise, Season-Wise Crop Production Statistics | Open Government Data (OGD) Platform India." *Open Government Data (OGD) Platform India*, Ministry of Electronics & Information Technology, Government of India, 21 Jan. 2022, data.gov.in/catalog/district-wise-season-wise-crop-production-statistics-0.

[16] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.