ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

International Journal of Mechanical Engineering

Forecasting Yield and Prices of Rice for Farmers in India

Aarti Karandikar* and Tanisha Agrawal

Department of Computer Science and Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur

Abstract:

Agriculture is the main occupation of India. Hence it is crucial to figure out the necessities that the agricultural sector needs. In this research paper, with the help of available information from the official government websites, different patterns have been identified, and a supervised machine learning model is also suggested. This paper also aims to offer ways to improve the accuracy of machine learning models. This will help the farmers to identify the factors that would increase the yield of crops(rice) and further estimate the price at which their crops will be sold.In turn helping the farmers to adjust the attributes with their availability getting more productivity and more profit. Python was used for the analysis of the different Machine Learning Algorithms - Multivariate Regression, Decision Tree, Gaussian Naive Bayes, SVM, etc, and it was found that Multivariate Regression gives the best and required result with an R square of 0.97 and 0.98 for yield and price calculation for respectively.

Keywords: Machine learning, Artificial Intelligence, Data Mining, Multivariate analysis, agriculture, etc

1. INTRODUCTION

Agriculture in India has struggled since independence.17% GDP of India is covered by agriculture.

Owing to the green revolution in the late 1960s, India showed great growth rates. The government was able to cater to food and money and bring about significant economic development. However, in the late 1990s, the situation started to decline, food limitation was observed, and farmers' income started to deteriorate.

To cater to the need for food, farmers started to use the land extensively for agriculture, which badly affected soil health. To improve the health of the soil, farmers started to use more fertilizers, which increased the cost price of cultivation of crops and led to the fact that fertilizers cannot provide all the nutrients to the soil. The present nutrient imbalance is seen for about 10 million tonnes per annum.

Table -1 : Requirement of rice

Commodity	Required production (million tonnes))
Rice	105

Objectives of research:

1. To identify the factors that would increase the yield of crops(rice) and further estimate the price at which their crops will be sold.In turn helping the farmers to adjust the attributes with their availability getting more productivity and more profit.
2. To explain the importance of agriculture in India.
3. To evaluate the current scenario of agriculture in India.
4. To study different machine learning models and suggest ways to improve their accuracy.
5. To suggest measures in which it can be improved.

Literature Review

Crop yield has been an important point of discussion in India owing to it's history from the green revolution. Many research has been done related to this. A recent study by Bhalla, G.S., and Hazell, P. 1997. in A Preliminary Exercise, Economic and Political Weekly December 27, A150-A154. Show the demand and need of food grains in 2020 (Bhalla and Hazell, 1997). The paper by (Gandhi et al. 2016) shows the experimental results obtained by applying SMO classifiers using the WEKA tool on the dataset of 27 districts of Maharashtra state, India. Temperature and precipitation however the fertilizers were not considered. Efficient Crop Yield Prediction in India using Machine Learning Techniques (Gulati ans Jha, 2020) here they have taken into consideration the different atmospheric conditions however irrigation which in further paper have proven to affect the yield greatly is not discussed. The research by Wu and their colleagues was done on SVM,Decision tree and KNN however Multivariate regression was not considered (Wu et al. 2009)

Most of the popular research for yield shows that SVM and decision trees to be the top choices for researchers. However when we consider prices not much research has been conducted here.

A recent research (Dhanapal et al 2021) showed that the decision tree was their preferred choice which in our case suffered from the problem of over-fitting. Lot of preprocessing was also required for this model making it less time efficient.

2. METHODOLOGY:

The data has been collected from various authorized sources mentioned in the references. It has been then processed and then split into train and test data sets, and then finally, a machine learning model has been developed.

2.1. Dataset

The data has been collected using authorized data repositories like www.data.gov.in. By analyzing the data, it has been found out that similar trends in the years have been observed for the attributes area under irrigation, fertilizers, rainfall, prices, yield, storage, and electricity.

It has been observed that even the attributes which were supposed to relate to yield are also related to prices because yield and prices are correlated.

We also found that rice yield is very much dependent on rainfall, irrigation, and storage and less on fertilizers.

Fig.1. Correlation matrix for electricity, storage, yield, prices, rainfall, fertilizers, and area under irrigation for rice

2.2. Preprocessing

The box plot diagram is very effective to show symmetry and outliers in data . It clarely defines the upper and lower quartiles of data. Fig.2.a clearly showed the outliers in the attributes rainfall,storage and fertilizers showed to have outlines which were removed by mean calculation method.

Based on this analysis of data, the data was processed, and the outliers were removed. Further pre-processing was done based on the requirement of the model.

Table 2 gives the overview of the processed data.

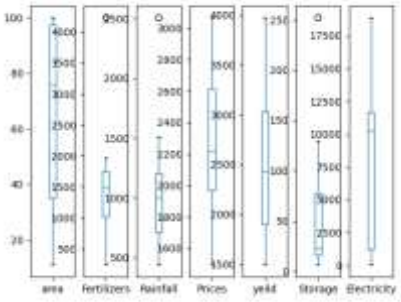


Fig.2.a. Box plox diagram

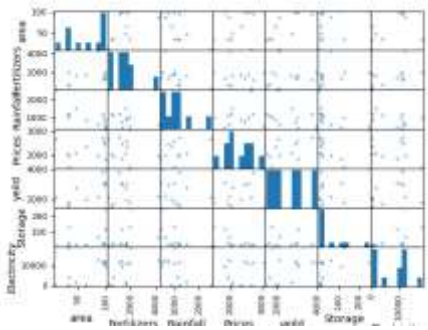


Fig.2.b. Histogram

Table.2. Processed data for the year 2015-2016(rice)

Area Under Irrigation (%)	Fertilizers (Thousand Tonnes)	Rainfall(mm)	Prices Rs. per (Quintel)	Storage Capacity (in Lakh Metric Tonnes)	Electricity consumption (GWh)	Yield - Kg./Hectare
47	1616	1507	2082	16.72	1524	2735
99.7	1943	512	3067	252	11513	3974
86.7	4230	612	1980	64.43	12671	2133
97	1698.15	987	2592	24.02	10970	3022
65	1696	874	2356	15	344	1948
94.4	1144.36	1204	2738	16.99	11548	3758
33.33	519	1201	2050	11.6	265	1491
98	1316.25	747	2471	20.88	11991	2933
11	242.62	2509	2677	6.29	34	2093
35	637.63	1136	1946	24.98	4025	1660
99.9	1347.4	437	1845	116.11	9506	3061
34.2	1966.54	1000	1495	129.66	18868	1752

2.3. Different Machine Learning methods

2.3.1. Multivariate Regression

This is a supervised ML technique which is an extension of multiple regression. It helps to overcome the flaws of multiple regression.It has one dependent variable and multiple

independent variables. It is given by the formula given in Fig 3.a. It helps to eradicate the flaws in multiple regression.

$$Y_i = \alpha + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_n x_i^{(n)}$$

Fig.3.a Multivariate equation

2.3.2. Decision tree ID3

It is also a supervised learning model and is based on a tree model for decisions for possible outcomes, utility and cost. It is also used for generating control statements. It is an entropy and information gain based model.It is built in a top-down manner.

2.3.3. Gaussian Naive Bayes

This model is a probabilistic based model. It uses the below formula.

$$P(C|K) = \frac{P(K|C) * P(C)}{P(K)}$$

It works well with small dataset and have the ability to provide uncertainty measurements.Gaussian processes are seen in infinite-dimensional generalization of multivariate normal distributions.It is based on normal distribution.

2.3.4. Support Vector Machine

It was Developed at AT&T Bell Laboratories and Vladimir Vapnik is known to be creator with colleagues (Boser et al., 1992, Guyon et al., 1993, Vapnik et al1997) It is robust prediction methods, It is based of statistics or VC theory proposed by Vapnik (1982, 1995) and Chervonenkis (1974). SVM training algorithm constructs a model which takes the responsibility to assign new examples to one category or others when training examples as marked as belonging to different categories,Hence it is a non-probabilistic binary linear classifier. SVM does mapping of training data to points in space to maximise width of gap between the two categories. New data are then mapped. They have the same space, they predict to belong in a category on the basis of gap

2.3.5. KNN

It was developed by Joseph Hodges and Evelyn Fix in 1951 It is a non-parametric classification method. It is based on closeness to other data points when given a distance, if close to one group of data points then they are grouped based on the parameter k

3. RESULTS:

Quantitative Analysis: The accuracy scores for the different models is as in Table 6. Accuracy score is a performance metric which is the ratio of true negatives and true positives to all positive and negative observations. It tells us about the number of times the model will give the correct results.

Table.6. Comparison table

Model	Accuracy for yield calculation in percentage	Accuracy for the price calculation in percentage
Multivariate Analysis	97	97
Decision Tree	100	100
Naive Bayes	60	33.33
KNN	60	33.33
SVM	100	60

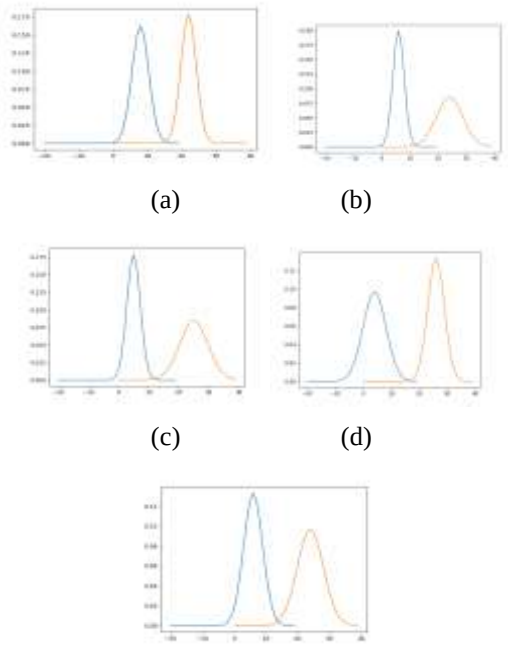
4. DISCUSSION

Most of the popular research for yield shows that SVM and decision trees to be the top choices for researchers. However when we consider prices not much research has been conducted here.

A recent research (Dhanapal et al 2021) showed that the decision tree was their preferred choice which in our case suffered from the problem of over-fitting. Lot of preprocessing was also required for this model making it less time efficient.

SVM worked well with yield but was less accurate with prices.With Multivariate regression it was observed that the more the number of attributes, the more is the accuracy of the model. Because the number of coefficients increases resulting in better value.

In the case of Gaussian Naive Bayes it was observed that the attributes that have distinct yes and no curves tend to give better results because they tend to have more area under yes/no curves, which means that only one attribute can provide better results.Fig 4 and 5 shows that Electricity gave best results due to distinct curves.

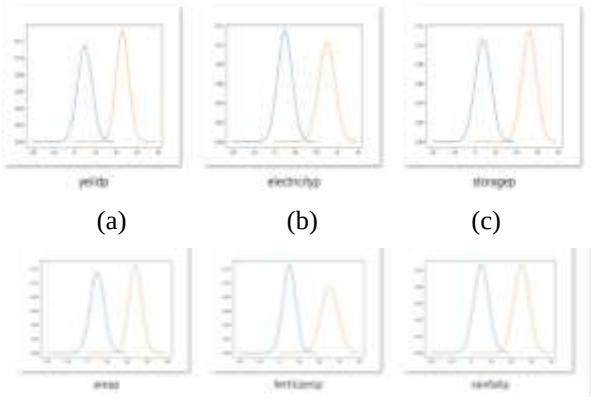


(a) (b) (c) (d) (e)

Fig4(a): area under irrigation **Fig4(b):**electricity **Fig4(c):**fertilizers **Fig4(d):** rainfall **Fig4(e):**storage

Blue → Yes, good yield probability

Orange→ no, not a good yield probability



(a) (b) (c)

(d) (e) (f)

Fig5(a): yield **Fig5(b):**electricity **Fig5(c):**storage
Fig5(d):area **Fig5(e):**fertilizers **Fig5(f):**rainfall

Blue → Yes, good price probability

Orange→ no, not a good price probability

5. CONCLUSION:

India needs agricultural reforms or strategies to improve agriculture. It has been evaluated that rainfall and irrigation are more required than fertilizers in rice cultivation. All factors contributing to yield also contribute towards prices since yield and prices are correlated. Similar patterns have been observed in the relation between attributes for the year 2012-2018. Various machine learning methods were evaluated to predict yield and prices, which can help farmers in planning. Table 6 gives a comparison of all the models. Based on the analysis, we can conclude that the Decision Tree and Multivariate analysis give the best results. However, the decision tree requires a lot of preprocessing and suffers from over-fitting . It was found that multivariate regression is the best-suited model.

With Multivariate regression it was observed that the more number of attributes, the more is the accuracy of the model. Because the number of coefficients increases resulting in better value. With Decision tree it was found that more categorized data is giving better accuracy. In case of Gaussian Naive Bayes it was observed that the attributes that have distinct yes and no curves tend to give better results because they tend to have more area under yes/no curves, which means that only one attribute can provide better results.

6. REFERENCES:

1. Bhalla, G.S., and Hazell, P. 1997. Foodgrains Demand in India to 2020: A Preliminary Exercise, Economic and Political Weekly December 27, A150-A154.
2. Datasets: www.data.gov.in, agriatglance (document by agriculture office)
3. Gandhi, N., Armstrong, L. J., Petkar, O., and Tripathy, A. K. 2016. Rice crop yield prediction in India using support vector machines In Proceedings of 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), KhonKaen, 1-5.
4. Gulati, P., Jha, S.K., 2020. Efficient Crop Yield Prediction in India using Machine Learning Techniques, International Journal of Engineering Research & Technology (IJERT), Vol. 8 – Issue 10.
5. Ranjani Dhanapal et al 2021 J. Phys.: Conf. Ser. 1916 012042
6. Tirpude, S., Karandikar, A., Welekar, R. 2020. An Approach for Environment Vitiation Analysis and Prediction Using Data Mining and Business Intelligence, 165, 327–338.
7. Veenadhari, S., Misra, B. and Singh, C. 2014. Machine learning approach for forecasting crop yield based on climatic parameters In Proceedings of International Conference on Computer Communication and Informatics, Coimbatore,1-5.
8. Wu, J., Olesnikova, A., Song, C. H., Lee, W. D. 2009. The Development and Application of Decision Tree for Agriculture Data IITSI, 16- 20.