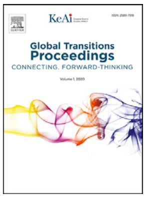
Contents lists available at ScienceDirect

Global Transitions Proceedings

journal homepage: http://www.keaipublishing.com/en/journals/global-transitions-proceedings/



Crop yield forecasting using data mining

Pallavi Kamath*, Pallavi Patil, Shrilatha S, Sushma, Sowmya S

Dept. Computer Science and Engineering, Shri Madhwa Vadiraja Institute of Technology and Management (Affiliated to VTU), Udupi 574115, India

ARTICLE INFO

Keywords:
Accuracy
Agriculture
Crop yield prediction
Data mining
Random forest algorithm

ABSTRACT

India is a heavily reliant on agriculture. Organic, economic, and seasonal factors all influence agricultural yield. Estimating agricultural production is a difficult task for our country, particularly given the current population situation. Crop production assumptions made far in advance can help farmers make the necessary planning for things like storing and marketing. Crop production prediction involves a huge amount of data, making it a perfect candidate for data mining methods. Data mining is method of accumulating previously unseen anticipated information from vast database. Data mining assists in the analysis of future patterns and character, enabling companies to make informed decisions. For a specific region, this research provides a fast inspection of agricultural yield forecast using the Random Forest approach.

1. Introduction

India's primary occupation is agriculture and the country's economy are entirely dependent on it for rural survival. Farming accounts for roughly 70% of the primary and secondary sectors. As a result, many farmers have begun to employ new technology and methods to improve their farming operations. People, on the other hand, are unaware of the importance of cultivating crops at the appropriate time and place. In this situation, using multiple elements that influence production to identify crop adaptability and yield can improve crop quality and yield, resulting in higher economic growth and profitability [2]. Crop development is a challenging phenomenon that agriculture input parameters recommend.

Data mining is method of accumulating previously unseen anticipated information from vast databases. Data mining assists in the analysis of future patterns and character, enabling companies to make informed decisions. The process of analysing, cleaning, and modelling data to generate useful knowledge and conclusions is known as data analysis [6].

Methods are used to convert the customer's raw data into valuable information. This research can be extended to agriculture as well. Most farmers relied on their long-term field experience with specific crops to forecast a greater yield in the coming season. Nonetheless, they do not receive a fair price for their crops. It typically occurs because of insufficient irrigation or poor crop selection, but it may also occur when crop yields are lower than expected. Due to a variety of factors, the farmers who make up the majority do not achieve the predicted Crop yield. That data set of crop yield which consists of many components. By studying the soil and atmosphere for the specific area, by which increase crop production, optimal crop can be estimated [10]. Advantage of our research mainly is Farmers will benefit from this forecast. To determine which crops are best for their farm based on soil type, ph., and fertilizer [11, 12].

2. Related works

Shailesh Shetty S *et al.* [1] This project supports farmers in evaluating which crop to grow in a specific area at a specific time and predicting whether it will be profitable or not. It gives the specifics by specifying whether the crop is profitable. As a result, this device aids farmers in their decision-making process, allowing them to save time.

Suvidha Jambekar *et al.* [2] Regression analysis is applied as a predictive modelling tool to predict crop production for crop production. The regression algorithms applied were, Multivariate Adaptive Regression Splines, and then Multiple Linear Regression, Random Forest Regression. According to the results, Random Forest Regression may be used to accurately estimate wheat, and rice, and maize production.

B. Devika, B. Ananthi *et al.* [3] Agriculture expands yield production to meet demand to limit overlapping, and the government encourages it for crop yield forecast on TamilNadu dataset imports. The regression method is put to the test of yield prediction capabilities in this study.

R. Vidhya *et al.* [4] They observed accuracy rate improves when a dataset with more features is used. As opposed to other approaches, such as Decision trees, linear regression, random forest algorithm is shown to be superior to other prediction algorithms. The included dataset incorporates a lot more variables resulting in more precise prediction.

Hetal Patel, Dharmendra Patel *et al.* [5] They measured performance of the classification algorithms Naive Bayes, J48, and Simple Cart. This crop prediction comparative analysis employs a large dataset and 10-

[*] Corresponding author.
E-mail addresses: pallakamath31@gmail.com (P. Kamath), sowmya.cs@sode-edu.in (S. S).

https://doi.org/10.1016/j.gltp.2021.08.008
Received 28 June 2021; Accepted 5 July 2021
Available online 13 August 2021

fold cross validation to give an indication of the predictive abilities of the data mining methods used.

Sangeeta, Shruthi G, *et al.* [6] They examined performance combining machine learning, Decision Tree and Polynomial Regression, Random Forest, and algorithms. Random Forest method outperforms the other algorithms in terms of yield prediction, according to the approach they have proposed. The Decision Tree model, like the random forest, polynomial regression, and decision tree models, classifies performance that shows changes in the dataset. As a result, we determined that the proposed model is more efficient than the current model for determining crop yield. The introduction of the above scheme will aid in the betterment of our country's agricultural practises. It can also be used to help farmers reduce their losses and increase crop yields in order to increase their resources in agriculture. To support our country's agricultural production, the system can be enhanced by merging it with other fields such as horticulture, sericulture, and others.

B A Harshanand, Swathi Sriram B Srishti, Chaitanya R, Kirubakaran Nithiya Soundari, V Mano Kumar, Varshitha Chennamsetti, Venkateshwaran G, Dr. Pramod Kumar Maurya, Akshay Prassanna S, *et al.* [7] determined the accuracy provided by the Decision Tree Model was about 92.66 percent. Maize, with a predicted increase of +3.32 percent, and Sunflower, with a predicted increase of 2.43 percent, are the top crop gainers. Niger will lose -7.81 percent of its crop, Moong will lose -4.2 percent of its crop, and Masoor will lose 2.4 percent of its crop. The new approach used to improve efficiency was the Random Forest Ensemble Method, which will be provided accuracy of about 97.57 percent, which is greater than Decision Tree Model. Rape is expected to gain +0.92 percent, while groundnut is expected to gain +0.445 percent. As a result, the comparison clearly demonstrates that ensemble approaches are often superior in terms of improving performance and efficiency while still providing high accuracy.

Saksham Garg, Parul Agrawal, Archit Agrawal, Aruvansh Nigam, *et al.* [8] In this paper, we looked upon ML algorithms which are used to determine harvests and based and mean absolute error techniques are compared. They converted three variables from the Indian government's official website, including temperature, rainfall, and production, into a final dataset. They considered 4 models: Logistic Regression XGBoost Classifier, KNN Classifier, Random Forest Classifier, and calculated accuracy of each model. They found that Random forest is best.

N.Rohit M.Vineeth and S.Bhanumathi, *et al.* [9] The proposed system predict the crop yield using random forest and deep learning model. And suggests the amount of fertilizer should be used for high yield. 2 different dataset -for crop data and fertilizer data are used. 80% used for training and 20% dataset is used for testing the data. The district, area, season, and production factors are used to build a machine learning model which predicts yield. By considering phosphorous, potassium, nitrogen amount in soil, quantity of fertilizer required is determined.

Kunal Teeda Nandini, Vallabhaneni, Dr.T.Sridevi *et al.* [10] Various models were discussed and their performance for a given dataset was measured in this paper, as well as artificial neural networks, Bayesian network, Cluster model, Conventional Methods for prediction using Decision trees, and ARIMA Prediction Model. K- Nearest Neighbours, multivariate Linear Regression prediction models are applied to analyse the rainfall and soil behaviour. And found that KNN classification model is the best because the accuracy of the training data is more compared to other.

3. Methodology

The overall Architecture of the proposed model using Random forest algorithm is described in Fig. 1.

The studies in this paper were carried out with PyCharm Community Edition 2021.1.1 × 64. The Important Classification Algorithm Random Forest is applied to the data collection provided from the Official Government website. To ensure accuracy, the datasets are examined. Random Forest is a supervised learning technique for classifying and predicting datasets. It will choose a collection of features at random from the dataset's attributes and construct a set of decision trees by locating the root nodes and splitting the attributes. Following the creation of the forest, the best decision is made based on the highest number of votes among the projected targets as the classifier's final prediction. Crop yield prediction systems provide for better planning and decision-making to increase production. The proposed system involves a prediction module based on data mining classification algorithm namely Random Forest used to forecast the yield of major crops based on historical data.

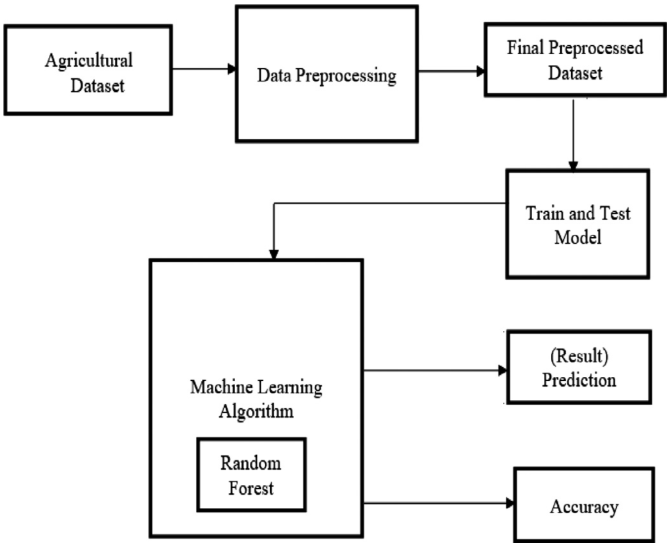


Fig 1. Block diagram of model

3.1. Agricultural dataset

Most of the research papers examined considered climatic variables such as, area, Temperature, Precipitation, and Humidity. Some soil agronomical parameters, such as chalky, clay, loamy, sandy, and so on, as well as different seasons, are included. The data of these variables were given as input. Initially dataset is collected which consisting of the parameters such as attributes like State Name, District name humidity, temperature, yield etc. Take into consideration any crops that will be planted in the region. This collected dataset is in csv format.

3.2. Pre-processing

A large dataset is needed for the of data mining application. The information gathered from different sources is often in raw form. It could include information that is incomplete, obsolete, or inconsistent. As a result, such redundant data should be filtered in this process. The information should be normalized. The provided data collection has many 'NA' values, which are filtered in Python.

Normalization is related to robust scaling, was also used but, uses the interquartile range instead of normalizing the data because the data set contains numeric data. Normalization reduces the size of the data by a factor of 0 to 1.

3.3. Train and test model

In the pre-processing step dataset will be divided into training dataset and testing dataset. This is the important step while creating model. The training dataset is used to train a model and testing dataset is used to evaluate the model. So, we fit the model with training data and test it with testing data.

3.4. Classification algorithm

Once data splitting is done next process is Creating and Training model using scikit-learn. The action of training machine learning model requires machine learning algorithm along with training data to grasp the pattern. Here we are using Random Forest algorithm which is well known supervised learning algorithm that works on bagging technique.

Random Forest Algorithm is a combination of number of decision tree. This algorithm is a classification algorithm based on ensemble classifier. It will divide the dataset into Training data and Testing data. Further training dataset is used to build the decision tree. Model will build a decision tree by considering training data and separates the weaker node from training data to get a better model. Each and individual training dataset will generate a decision tree and then generate random forest. The general idea of the bagging method is that an aggregate of mastering output will increase the overall result. Random forest algorithm builds multiple decision trees during training. Predictions made from these decision trees will be collected and the final output will be the one which is having maximum votes. Jupiter notebook is a platform which used to create the trained model using Random Forest Algorithm.

Algorithm of Random Forest:

Input:

```
i) node from the decision tree, if node,
attribute = k then the split is done on the Kth
attribute
ii) V as the value obtained from the decision
tree then Vk= the value of kth attribute
Output:
label of V
If node is a Leaf, then
Return the value predicted by d
Else
Let k= node. Attribute
If j categorical then
Let v= Vk
Let Cv = child node corresponding to the
attribute's value v
Return Classify (Cv, V)
Else K is real valued
Let t= node. threshold (split threshold)
If Vk<t then
Let Cl = child node corresponding to (<t)
Return Classify (Cl, V)
Else
Let Ch = child node corresponding to (>=1)
Return Classify (Ch, V)
```

3.5. Predict yield

The trained model is used to predict the output on new input. Here we saved the trained model in a file so that model can be predict on the new input.

In this system we used pickle format developed in Jupyter, to store the trained Machine Learning model which stores the object in binary stream and evaluate the model with testing dataset. This prediction model contains random forest algorithm that learn properties from training data by using data it will make the predictions.

3.6. Accuracy

Accuracy is the one of the metrics uses for evaluating classification model. Accuracy is calculated by dividing number of correct predictions by total number of predictions.

$$Accuracy = \frac{Numberofcorrectpredictions}{Totalnumberof\ Predictions} \quad (1)$$

Table 1
Comparison of different Models wrt Accuracy [8]

MODEL	ACCURACY (in percentage)
Random Forest Classifier	67.80
XGBoost Classifier	63.63
KNN Classifier	43.25
Logistic Regression	25.81

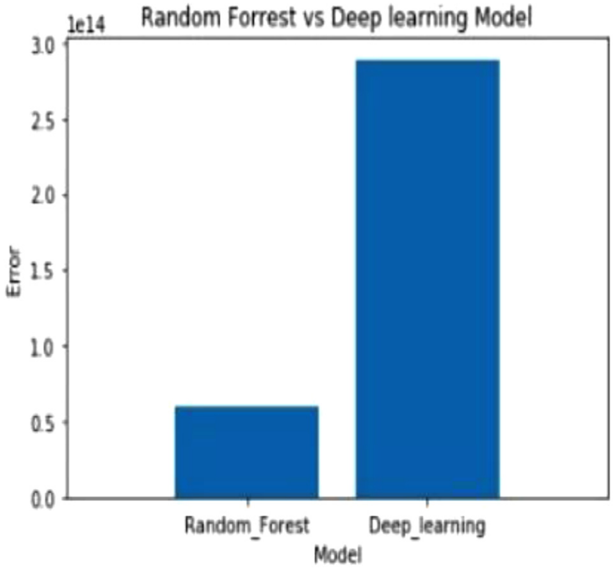


Fig 2. Comparison between random forest and deep learning [9]

We have achieved the 98% accuracy which means this model is good for predicting yield.

Comparison with different model:

By considering the different algorithm while predicting the yield, The Random Forest Algorithm achieved High Accuracy. This is because the Random forest will construct the decision tree for individual set of training dataset and then combine the multiple decision tree into to a single decision tree and it will predict the yield by considering the Average value of the Tree [8]. Analysis of different algorithm is considered in Table 1 with respect to accuracy and comparison between Random forest and Deep learning model [9] is graphically explained in Fig 2.

4. Result and discussion

In this paper effort is made in order to know the region-specific crop yield analysis and it is processed by implementing by random forest algorithm. In this project have chosen dataset which in .csv format. For the training purpose 80% of data is used and remaining 20% of data is used for testing. After the successful training and testing next step is finding the accuracy of the model. We have achieved a good accuracy which means this model is good for predicting yield. We have designed the Website which consists of Four Functional Modules as shown in the Fig. 3.

1) Crop Module: This module will provide the list of available crops. On selection of each one of it will give the detailed description of the crop.
2) Soil Module: This module will provide the list of available soils. On selection of each one of it will give the detailed description of the soil.
3) Weather Module: In this module by entering the city name the user can get the live weather forecast. Openweatherapi is free open source weather data. By using weather API key can fetch the current or historical weather data.

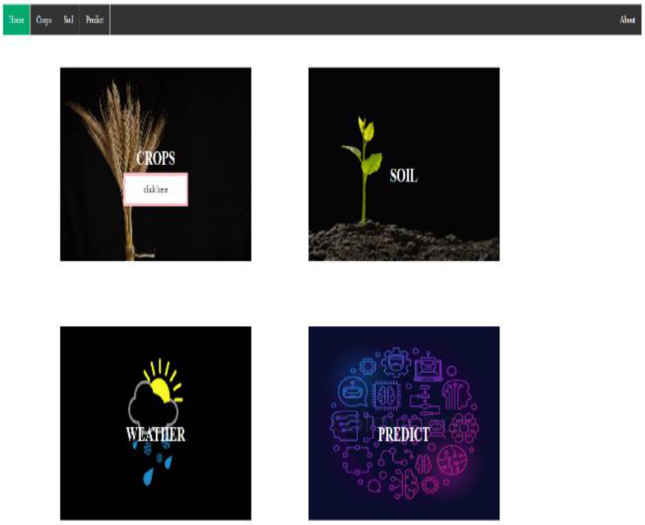


Fig 3. Module description

Table 2
Comparing the accuracy

Model	Accuracy
Proposed Model	98%
Saksham Garg *et al.* *(2019)*	67.80%
Shriya Sahu *et al.* *(2017)*	91.43%

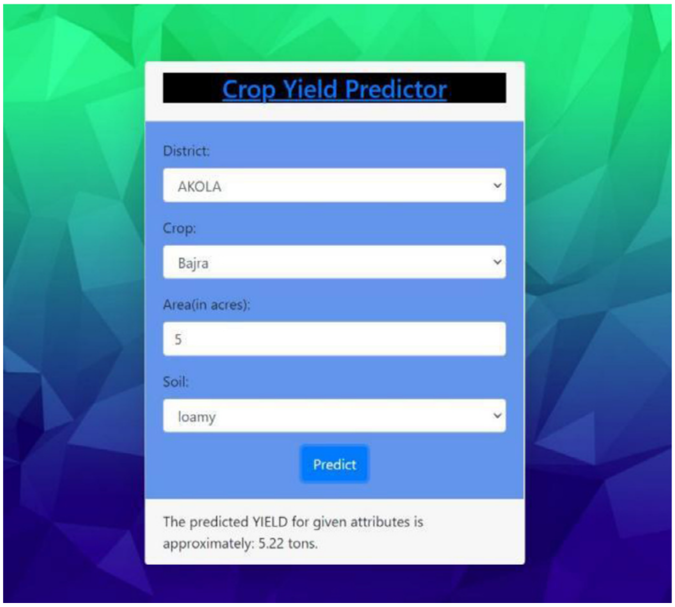
4) Predict: This predict module allows the user to select the district name, crop name, soil type and area. After selecting these values user can click the predict button to get the estimated yield [15-17].

Comparative analysis of Random forest algorithm accuracy [8, 13, 14] is mentioned in Table 2

Fig. 4 Webpage defines the yield (tons) predicted by the consumer.

Data visualization is done by plotting the Yield variable with different parameter

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make easier for the human to grasp and pull insights. The main goal of data visualization is to make it easier to identify patterns, trends, and outliers in large data sets. Fig. 8 gives the proper idea about the information that is present in the dataset [18-20].



Crop Yield Predictor

District:
AKOLA

Crop:
Bajra

Area(in acres):
5

Soil:
loamy

The predicted YIELD for given attributes is approximately: 5.22 tons.

Fig 4. Prediction module

Pair-plot is one which is used for visualization of dataset which is graphically represented in Fig. 5. This is the module of seaborn library. From the image below we can observe variation in each plot [21-22].

Jointplot is also a seaborn library which is used to quickly visualize the relationship between two variables which is graphically shown in Figs. 6 and 7. In the below figure it gives relationship between yield and year.

Fig. 9 shows the graphically comparison between Actual and predicted value of allocated dataset.

5. Conclusion

The paper discussed machine learning algorithms for predicting crop yield based on temperature, season, and location. A Yield prediction for

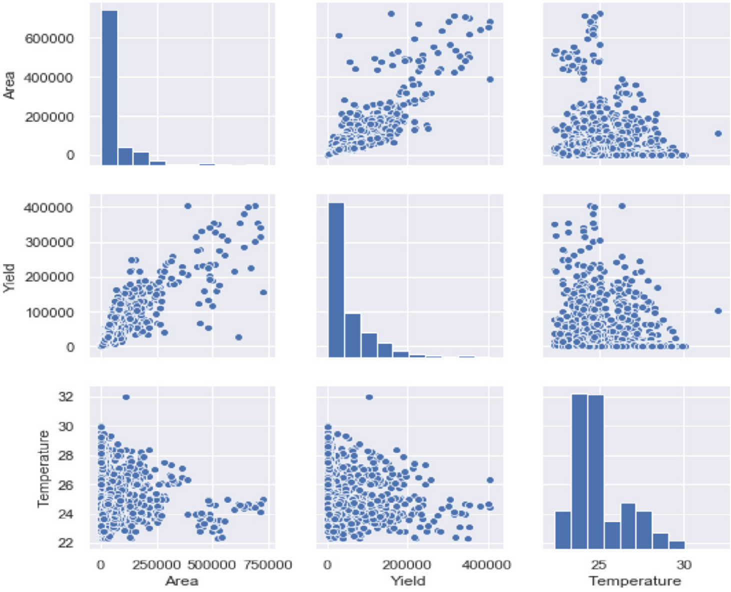


Fig 5. Pairplot between area, yield and temperature

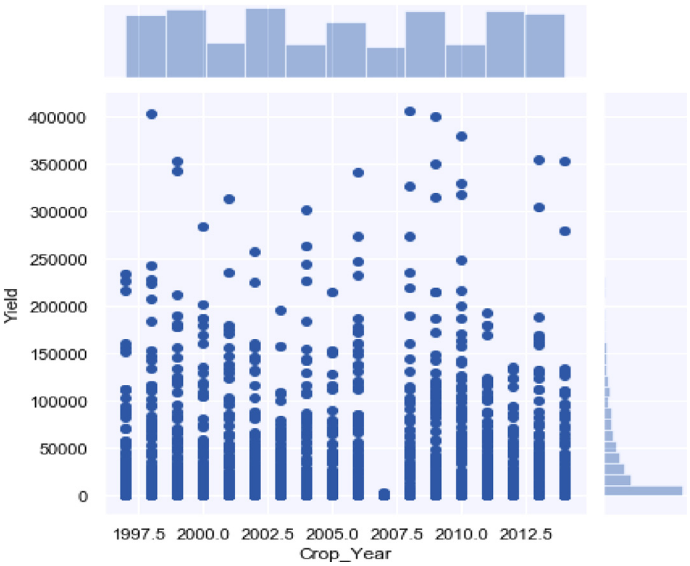


Fig 6. Jointplot of yield vs year

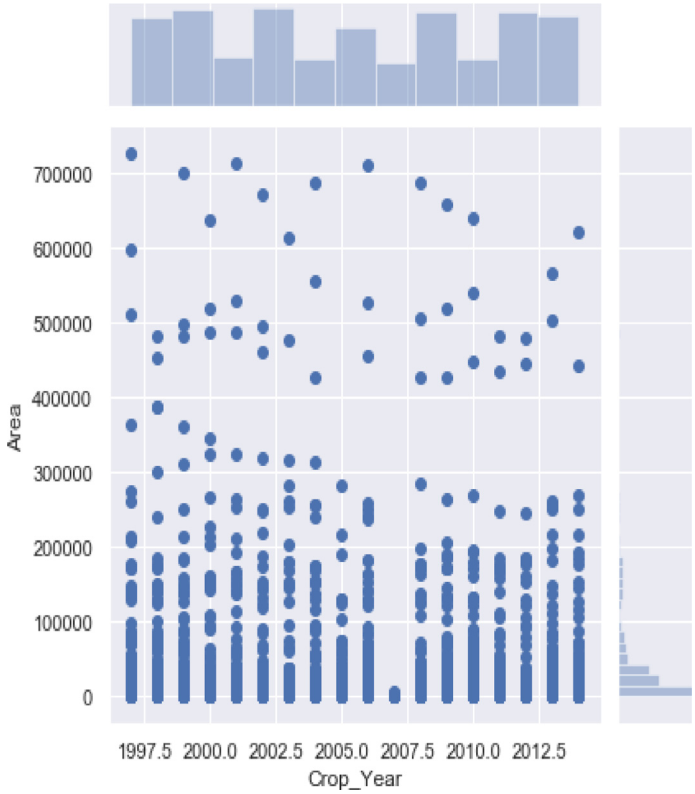


Fig 7. Jointplot of area and year

a specific district can be made by combining Precipitation, Temperature, and other parameters such as season and location. When all the factors are considered, Random Forest emerges as the greatest classifier. The dataset which is in use with more features increases the accuracy rate. Random forest is the superior prediction algorithm when compared to other technologies that are multiple linear regression and decision trees. Our dataset contains a lot more variables, resulting in more accurate predictions. The introduction of this project which are helpful to the farmers to reduce their losses and increase crop yields to increase their resources in agriculture. This will not only help farmers choose the best crop to cultivate in the future season, but it will also help bridge the

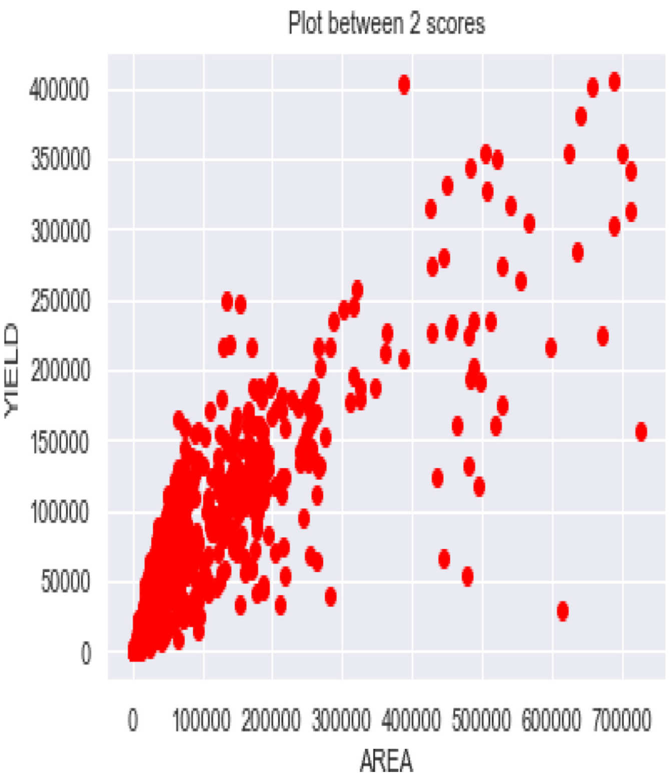


Fig 8. Data visualization between yield and area

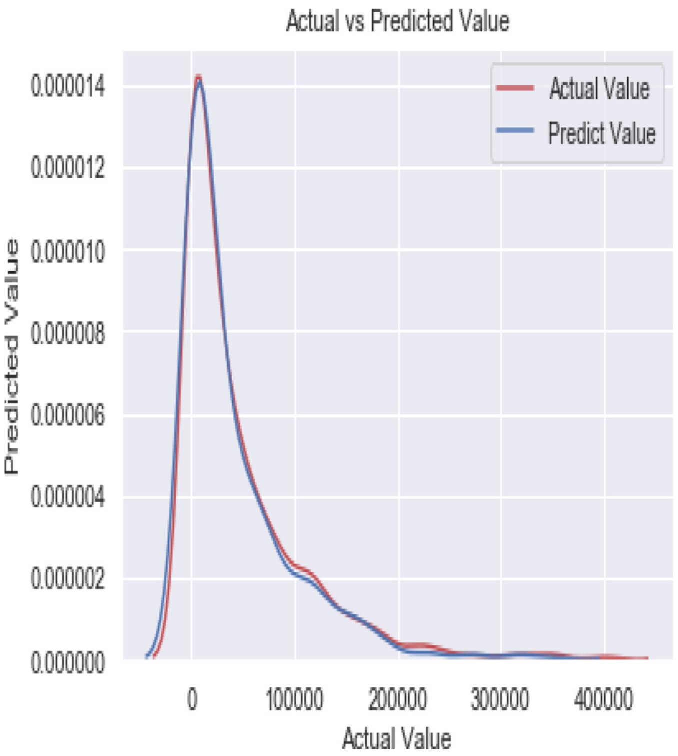


Fig 9. Comparison between actual and predicted values

technological and agricultural divide. Limitation of our project is, Yield is predicted for 100acres and implemented for 30 districts. The Future work of our project is to overcome our limitations.

References

[1] Shailesh Shetty S, Akshatha, Anet P James, Chaitra M Poojary "Crop analysis and profit prediction using data mining techniques" (IJERT).

[2] Shruthi G Sangeeta, Design and implementation of crop yield prediction model in agriculture, Int. J. Sci. Technol. Res. 8 (01) (JANUARY 2020).

[3] B. Devika, B. Ananthi, Analysis of crop yield prediction using data mining technique to predict annual yield of major crops, Int. Res. J. Eng. Technol. (IRJET) 05 (12) (Dec 2018).

[4] R. Vidhya, Pragya Mathur, Shivani Sai Valluri, Crop yield prediction using random forest, Int. J. Adv. Sci. Technol. 29 (9s) (2020) 3084–3086.

[5] Hetal Patel, Dharmendra Patel, "A comparative study on various data mining algorithms with special reference to crop yield prediction." Indian J. Sci. Technol.

[6] Suvidha Jambekar, Shikha Nema, Zia Saquib, "Prediction of Crop Production in India Using Data Mining Techniques", IEEE, 2018 978-1-5386-5257- 2/18/$31.00 ©.

[7] BA Harshanand, Swathi Sriram B Srishti, Chaitanya R, Kirubakaran Nithiya Soundari, V Mano Kumar, Varshitha Chennamsetti, Venkateshwaran G, Dr. Pramod Kumar Maurya, Akshay Prassanna S, "Crop value forecasting using decision tree regressor and models" Eur. J. Mol. Clin.l Med.

[8] Saksham Garg, Parul Agrawal, Archit Agrawal, Aruvansh Nigam, "Crop yield prediction using machine learning algorithms" 2019 Fifth International Conference on Image P rocessing(ICIIP).

[9] N. Rohit, M. Vineeth, S. Bhanumathi, Crop yield prediction and efficient use of fertilizers, International Conference on Communication and Signal Processing, April 4-6, India, 2019.

[10] Kunal Teeda Nandini, Vallabhaneni, T. Sridevi "Comparative analysis of data mining models for crop yield by using rainfall and soil attributes" Proceedings of the 2nd (ICICCT 2018) IEEE Xplore Compliant.

[11] B.Jabber Potnuru Sai Nishant, Venkat, Bollu Lakshmi Avinash, Pinapa Sai, Crop Yield Prediction based on Indian Agriculture using Machine Learning, INCET, Belgaum, India, Jun 5-7, 2020.

[12] Yogesh Gandge, Sandhya "A study on various data mining techniques for crop yield prediction" 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT).

[13] Shriya Sahu, Meenu Chawla "An efficient analysis of crop yield prediction using hadoop framework based on random forest approach" International Conference on Computing, Communication and Automation (ICCCA2017).

[14] Dr. R. Sujatha, P.Isakki Devi, A Study on Crop Yield Forecasting Using Classification Techniques IEEE, 2016. 978-1-4673-8437-7/16/$31.00 ©.

[15] S.I. Chu, C.L. Wu, T.N. Nguyen, B.H. Liu, Polynomial computation using unipolar stochastic logic and correlation technique, IEEE Trans. Comput. (2021).

[16] T.N. Nguyen, V.V. Le, S.I. Chu, B.H. Liu, Y.C. Hsu, Secure localization algorithms against localization attacks in wireless sensor networks, Wireless Pers. Commun. (2021) 1–26.

[17] S. Veenadhari, Dr. Bharat Misra, Dr. CD Singh "Data mining techniques for predicting crop productivity – a review article"

[18] D.N. Tran, T.N. Nguyen, P.C.P. Khanh, D.T. Trana, An iot-based design using accelerometers in animal behavior recognition systems, IEEE Sens. J. (2021).

[19] P. Subramani, G.B. Rajendran, J. Sengupta, R. Pérez de Prado, P.B. Divakarachari, A block bi-diagonalization-based pre-coding for indoor multiple-input-multiple-output-visible light communication system, Energies 13 (13) (2020) 3466.

[20] V. Rajeswari, K. Arunesh, Analysing soil data using data mining classification techniques, Indian J. Sci. Technol. 9 (19) (May 2016), doi:10.17485/ijst/2016/v9i19/93873.

[21] L.J.L. Sujan, V.D. Telagadi, C.G. Raghavendra, B.M.J. Srujan, R.V. Prasad, B.D. Parameshachari, K.L. Hemalatha, Joint reduction of sidelobe and pmepr in multicarrier radar signal, in: Cognitive Informatics and Soft Computing, Springer, Singapore, 2021, pp. 457–464.

[22] Rajendran, G.B., Kumarasamy, U.M., Zarro, C., Divakarachari, P.B. and Ullo, S.L., 2020. Land-use and land-cover classification using a human group-based particle swarm optimization