Comparison of Various Classification Techniques for Prediction of the Agriculture Production Based on Different Parameters Rainfall, Temperature

[1] Manpreet Kaur Research Scholar, Department of Computer Applications, Guru Kashi University, Talwandi Sabo, PB, India

[2] Dr. Dinesh Kumar Associate Professor, Department of CSE, Guru Kashi University, Talwandi Sabo, PB, India

Abstract: *The classification techniques based on various machine learning techniques are having use for the Big data analysis. This will be useful in identifying the classification and then finally the prediction which will be useful for the decision managers for having quality decisions. There are various types of supervised and unsupervised learning techniques which are having capabilities in the terms of driving the analysis. This analysis will be useful for having identification of relationship between the various attributes which is required to device the analysis. There are various supervised learning techniques which are useful to drive the analysis. These techniques are SVM, Logistic regression, KNN, Naïve Bayes, Tree, Neural network. The relative comparison of this technique is done in the terms of various parameters for example AUC, CA, F1, Recall and precision. The accuracy in the terms of AUC, CA is highest for the Naïve Bayes. This shows the Naïve Bayes is having higher true positives, true negative ratio. The proposed technique is having higher accuracy of 81% which is far above than all the remaining techniques. The confusion matrix for the Naïve Bayes is having true positive count as 729, true negative at 103. This shows that the true positive and true negative count is far above for this technique compared to the other techniques.*

Keywords: *Classification, Agriculture yield, Prediction, Accuracy*

1. BACKGROUND

As in current situation India is the second largest producer of the farm based products in whole world. In India almost 70% of the population is directly dependent on the agriculture production. The agriculture is one of the broadest economic sector and affects the socio economic culture of the country. The agriculture is a unique business which has various factors which affect its

growth or we can say output[12][13]. The climate based factors are main focusing points which have greater impact on the agriculture production. Other than the rainfall there are various parameters for example soil, climate, cultivation, irrigation, fertilizers, temperature, rainfall, harvesting, pesticide weeds and other factors. The crop yield prediction will be considered as the main focusing point for the companies involved in the processing of the agriculture produce. The supply chain management is the core issue which requires some capacity building and then utilization if the correct level prediction is done. There are various other businesses that are directly be connected to the agriculture for example seed, fertilizer, agrochemical and agricultural machinery industries plan production and marketing activities etc. which requires quite high number of prediction based activities so that right prediction can be done. The government on the other hand is involved in the prediction scenario so that policy framework can be built[10].

- The policy decisions taken by the government will be having higher level of futuristic prediction scenario.
- It will help in collecting the legacy data related to the agriculture produce and its most effecting factors.

The government broader level insurance scenario can also be build such that premium amount can be predicted by having correct estimate of the agriculture produce.

Data mining is new field in the sector of agriculture, because previously there was no such system which is used for the prediction of the agriculture produce based on various relevant factors. The level of the effect of these parameters on the agriculture prediction can be identified. The data mining field in integration to the machine learning techniques became a powerful

process to identify the prediction accuracy. There are two types of machine learning techniques, one is in the category of supervised and other is the unsupervised learning techniques. In the supervised learning techniques, the set of rules are mentioned by the researcher based on the parameters or attributes values. These conditions are adjusted based on the variation in the data so that classification and then the prediction accuracy can be enhanced. There are various techniques which comes in the supervised learning process are based on the classification process, these techniques are KNN, Random Forest, Decision Tree, Naïve Bayes etc. These techniques classify the data based on various attributes values so that given dataset can be proportionally classified into different classes, and then finally the prediction perspective is done which will be used for the prediction based on the target variable. The unsupervised learning technique is another way of for identifying the relation between different attributes. The data belongs to different attributes are having natural relationship such that right perspective of relationship can be developed. The regression based analysis is the right technique for those datasets having various attributes which are having natural relationship for example age and salary [1].

2. LITERATURE SURVEY

[1]P. S. Budi Cahyo Suryo et al.(2019): The smart agriculture is the need of the hour, as various applications are being developed to automate the decisions regarding various angles, the large dataset is being used for the prediction perspective, LSTM based technique is used for reduction in the RMSE value compared to the back propagation process. The researcher has conducted different aspects for checking whether the back propagation is having superior results. But the LSTM based technique provides better results compared to the back propagation. The proposed

technique is having RMSE value at 0.8 compared to it, the back propagation is having 0.1 results.

[2]A. Savla et al.(2015): The machine learning is highly suitable process for having various prediction in various different fields. The agriculture is new entrant into this, the yield prediction of soybean crop is main focusing issue, so that the decision regarding different aspects of agriculture production can be taken place. The researcher in current paper has put various prediction based techniques based on machine learning in the supervised learning techniques. The proposed mechanism based on the large dataset related to the agriculture production will be suitable for handling large block of data. The comparative analysis of different techniques on the basis of different parameters for example accuracy.

[3]P. S. Nishant et al.(2020):The author in this paper has given the research on the prediction of agriculture produce for various different crops planted in Indian context. The researcher has provided the technique based on advanced regression techniques like Kernel Ridge, Lasso and ENet algorithms to predict the yield and uses the concept of Stacking Regression for enhancing the algorithms to give a better prediction. The proposed technique uses various parameters based on state, district, season and the area so that various different previous years data and its impact with respect to the temperature and humidity can be considered. The proposed technique shows the best of the results compared to other techniques available for the prediction for the agriculture produce.

[4]Meizir and B. Rikumahu(2019): author in this paper has given the proposal for prediction of stock prices and the agriculture produce. In both the areas there are higher level of variations in the stock prediction accuracy because there are various factors which will affecting the stock

prices, these factors are hard to prediction and are having higher level effect the stock prices. The technical analysis with Artificial Neural Network Backpropagation method is used by the researcher for the prediction scenario. There are various factors which are being considered to compare the performance of the proposed technique with the base technique. The proposed technique is having lower level variations for the prediction compared to the actual. The predicted values of Artificial Neural Network Backpropagation them means showing a promising result.

[5]K. Tripathy *et al*.(2011):The author in this paper has proposed a technique for the pest/disease management. The author in this paper has taken the dataset related to the dynamic crop data over to the different years. There are various proposed mechanisms that suitable for the prediction of the results. The large sensor devices are being placed at different positions to collect the data related to the plants in specific area, so that sensory data can be analyzed to interpret the effect of various factors on the plant disease and yield. The study will help in generating the analysis about various aspects and will provide the way for accurate prediction of the disease and its management practices over the years.

[6]S. Nagini et al.(2016):The researcher in this paper has proposed a research on the prediction of the agriculture produce based on the past records. In India where there are large population which is dependent on the agriculture if will get good prediction then the level of the risks in the agriculture produce can be minimized. There are plenty of the factors which affect the crop yield, these parameters are Water, Nitrogen, Weather, Soil characteristics, Crop rotation, Soil moisture, Surface temperature and Rain water etc. which are having direct relevance in context to the production. The proposed methodology provides the good way for the prediction so that with the knowingness of various parameters expected flow of the whole scenario can be predicted. There

are various critical parameters which are having direct impact on the agriculture produce, this parameter is mainly rainfall. The author has given the research in the line of regression models like Linear, Multiple Linear, Non-linear models for identifying the high accuracy and also the relative accuracy of these techniques can be compared so that best technique whose accuracy is high can identified and adopted.

3. METHODOLOGY

The methodology is based on the sequential steps that are taken for identification of the prediction for the agriculture yield. There are various machine learning techniques that are to be tested on the give dataset on the basis of accuracy of the prediction. The confusion matrix is identified by comparing the evaluated prediction with the actual, four different parameters are set, one is the true positive, second is true negative, third is false positive, fourth is false negative.

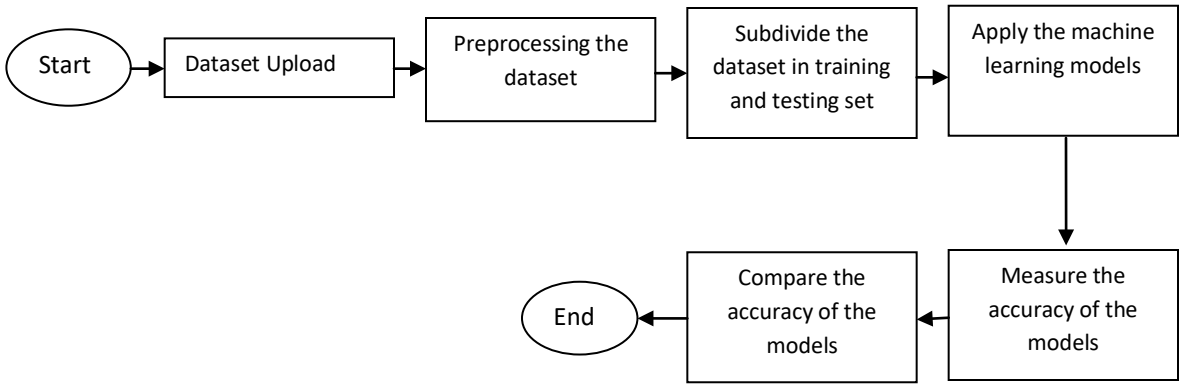

Fig. 1 Methodology

4. DATASET

The dataset based on the traditional data related to various factors for example temperature, humidity, soil conditions and various other attributes which either affect the agriculture produce directly or may be indirectly. The dataset is available at Kaggle as open source freely available

dataset can be used for prediction of the agriculture production or yield with results of different values of parameters.

5. TRAINING AND TESTING SET

The whole dataset is subdivided into two parts for applicable on the machine learning techniques for the prediction scenario. The first set is the training set and second set is the testing set. The 70% is kept for the training set and 30% is set for the testing set. The selected model is made to learn from the training set, such that learned things are applied on the testing set. The higher will be the learning higher will be accuracy on the testing set.

6. RESULTS

There are various classifiers that are used for classification of the dataset based on the agriculture production and its various factors which affect the agriculture yield. The supervised learning techniques for classification of agriculture data are having different levels of accuracies. These accuracies are measures in the terms of four factors, true positive, true negative, false positive, false negative.

6.1 SVM based classification

	Predicted		
Actual	741	66	807
	134	69	203
	875	135	1010

Table 1 confusion matrix using SVM

The table 1 represents the four variables, true positive, true negatives, false positive, false negative. The SVM based classification for agriculture data is having higher accuracy.

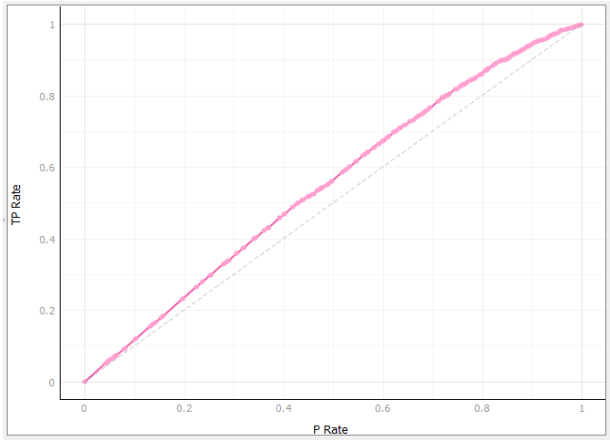

Fig.2 Lift chart result

The graph in fig. 2 shows the TP rate and positive rate, the true positive count is higher compared to the false negative and false positive.

6.2 Tree based classification

	Predicted		
Actual	717	90	807
	114	89	203
	831	179	1010

Table 2 confusion matrix

The graph shows the confusion matrix for the tree based classification. The true positive and true negative rate for the Tree based technique is higher compare to the false negative, false positive.

6.2.1 Parameters results.

Model	AUC	CA	F1	Precision	Recall
Tree	0.646	0.798	0.793	0.789	0.798

Fig. 3 Parameter result

The figure 3 shows the parameters results with respect to AUC, CA, F1, Precision, Recall.

6.2.2 Lift chart results

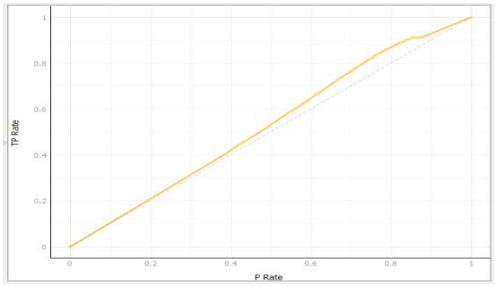

Fig. 3 Lift chart

The graph in fig. 2 shows the TP rate and positive rate, the true positive count is higher compared to the false negative and false positive.

6.3 KNN based classification

	Predicted		
Actual	775	32	807
	162	41	203
	937	73	1010

Table 3 Confusion matrix for KNN

Table 3 shows the confusion matrix for the agriculture produce data based on KNN based classification.

6.3.1 Parameters results

Model	AUC	CA	F1	Precision	Recall
kNN	0.715	0.808	0.770	0.774	0.808

Fig. 4 Parameters values

Fig. 4 shows the results for different parameters based on KNN based classification.

6.3.2 Life chart results

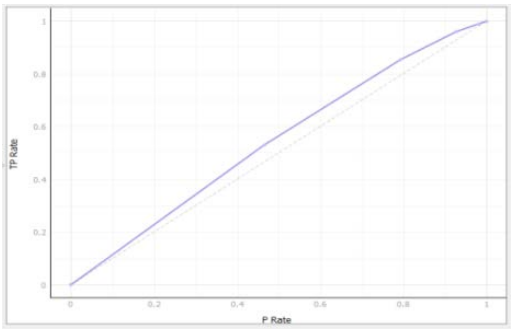

Fig. 5 Lift chart

The graph in fig. 5 shows the TP rate and positive rate, the true positive count is higher compared to the false negative and false positive.

6.4 Logistic regression based classification

	Predicted		
Actual	771	36	807
	125	78	203
	896	114	1010

Table 4 confusion matrix based on logistic regression.

6.4.1 Parameters results

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.715	0.808	0.770	0.774	0.808
Logistic Regression	0.799	0.841	0.822	0.825	0.841

Fig. 6 Parameters results

The fig. 6 shows the parameters results based on different parameters.

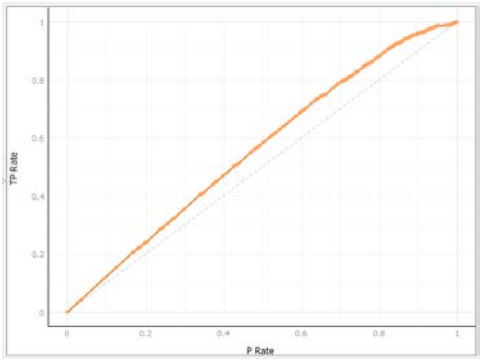

Fig. 7 Lift chart

The graph in fig. 7 shows the TP rate and positive rate, the true positive count is higher compared to the false negative and false positive.

6.5 Naïve Bayes based classification

	Predicted		
Actual	729	78	807
	100	103	203
	829	181	1010

Table 5 confusion matrix based on Naïve bayes

6.5.1 Parameters results

Model	AUC	CA	F1	Precision	Recall
kNN	0.715	0.808	0.770	0.774	0.808
Logistic Regression	0.799	0.841	0.822	0.825	0.841
Naive Bayes	0.810	0.824	0.820	0.817	0.824

Fig. 8 Parameters results

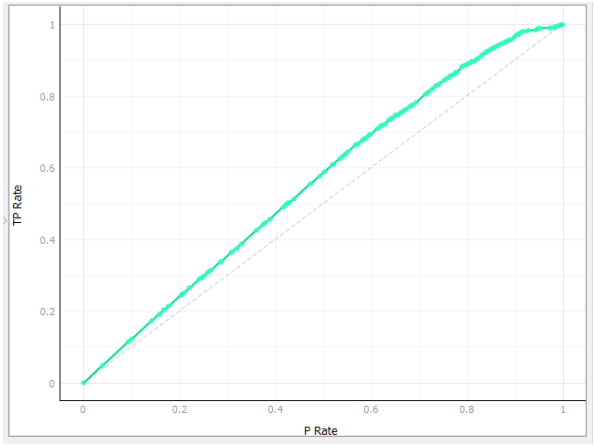

Fig. 9 Lift chart

The graph in fig. 9 shows the TP rate and positive rate, the true positive count is higher compared to the false negative and false positive.

6.6 Neural network based classification

	Predicted		
Actual	733	74	807
	122	81	203
	855	155	1010

Table 6 confusion matrix

6.6.1 Parameters results

Model	AUC	CA	F1	Precision	Recall
kNN	0.715	0.808	0.770	0.774	0.808
Logistic Regression	0.799	0.841	0.822	0.825	0.841
Naive Bayes	0.810	0.824	0.820	0.817	0.824
Neural Network	0.748	0.806	0.796	0.790	0.806

Fig. 10 parameters results

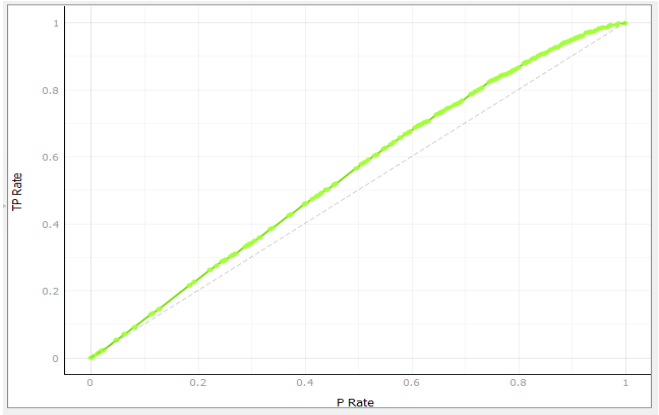

Fig. 11 lift chart

7. COMPARATIVE TABLE

	AUC	CA	F1	Precision	Recall
Tree based classification	0.646	0.798	0.793	0.789	0.798
KNN	0.715	0.808	0.770	0.774	0.808
Logistic regression	0.799	0.841	0.822	0.825	0.741
Naïve Bayes	0.810	0.824	0.820	0.817	0.824
Neural network	0.748	0.806	0.796	0.790	0.806
SVM	0.748	0.802	0.786	0.779	0.802

Table 7 Comparison for all the parameters for different techniques

The table shows the comparative accuracy for all the classification techniques, there are various parameters which are used for comparing all the techniques. The Naïve Bayes is the best suitable technique as far as accuracy is concerned. The classification and then finally prediction based on Naive Bayes is having highest accuracy.

8. CONCLUSION

The classification based on the machine learning techniques is the emerging field for different areas where there is some sort of quality in the decision making is required. The health sector is new entrant to the data analytics. This provides the way for early prediction for any type of disease based on the different parameters. The classification and prediction for the business decision making is used by various levels of the decision managers for enhancing the quality of the decisions. There are abundance numbers of techniques and tools are available which provide the way to achieve the higher accuracy for the prediction. There are abundance numbers of machine learning techniques which provide the way with different levels of accuracy. The accuracy using Naïve Bayes is highest out of all the other technique. The Naïve Bayes is having accuracy at 81%. This is far above than the other classification and prediction techniques.

9. FUTURE WORK

The classification techniques based on the supervised learning is the best way for the classification and the then prediction for specific aspect related to the different fields. There are various algorithms based on the machine learning are suitable for prediction with higher level of accuracy. The proposed algorithm named as the Naïve Bayes is having higher accuracy compared to various algorithms. The supervised machine learning algorithm is having higher positivity rate compared to the false rate. The accuracy can be tested with the adjustment in the training and testing set. The higher training set is having higher accuracy.

REFERENCES

1. P. S. Budi Cahyo Suryo, I. Wayan Mustika, O. Wahyunggoro and H. S. Wasisto, "Improved Time Series Prediction Using LSTM Neural Network for Smart Agriculture Application," 2019 5th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2019, pp. 1-4, doi: 10.1109/ICST47872.2019.9166401.

2. A. Savla, N. Israni, P. Dhawan, A. Mandholia, H. Bhadada and S. Bhardwaj, "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture," 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore, India, 2015, pp. 1-7, doi: 10.1109/ICIIECS.2015.7193120.

3. P. S. Nishant, P. Sai Venkat, B. L. Avinash and B. Jabber, "Crop Yield Prediction based on Indian Agriculture using Machine Learning," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154036.

4. Meizir and B. Rikumahu, "Prediction of Agriculture and Mining Stock Value Listed in Kompas100 Index Using Artificial Neural Network Backpropagation," 2019 7th International Conference on Information and Communication Technology (ICoICT), Kuala Lumpur, Malaysia, 2019, pp. 1-5, doi: 10.1109/ICoICT.2019.8835284.

5. A. K. Tripathy et al., "Data mining and wireless sensor network for agriculture pest/disease predictions," 2011 World Congress on Information and Communication Technologies, Mumbai, India, 2011, pp. 1229-1234, doi: 10.1109/WICT.2011.6141424.

6. S. Nagini, T. V. R. Kanth and B. V. Kiranmayee, "Agriculture yield prediction using predictive analytic techniques," 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida, 2016, pp. 783-788, doi: 10.1109/IC3I.2016.7918789.

7. A. H. Manek and P. K. Singh, "Comparative study of neural network architectures for rainfall prediction," 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), Chennai, India, 2016, pp. 171-174, doi: 10.1109/TIAR.2016.7801233.

8. Y. Jun, C. Weiwei, L. Yu, H. Jinmin, D. Jiannan and L. Wenjie, "Grey Relevant Analysis and Prediction on Agriculture Mechanization of China," 2011 Fourth International Conference

on Intelligent Computation Technology and Automation, Shenzhen, China, 2011, pp. 97-100, doi: 10.1109/ICICTA.2011.315.

9. A. Vohra, N. Pandey and S. K. Khatri, "Decision Making Support System for Prediction of Prices in Agricultural Commodity," 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 2019, pp. 345-348, doi: 10.1109/AICAI.2019.8701273.

10. Z. Chen et al., "Charms - China Agricultural Remote Sensing Monitoring System," 2011 IEEE International Geoscience and Remote Sensing Symposium, Vancouver, BC, Canada, 2011, pp. 3530-3533, doi: 10.1109/IGARSS.2011.6049983.

11. L. N. de Castro, F. J. von Zuben and W. Martins, "Hybrid and constructive neural networks applied to a prediction problem in agriculture," 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227), Anchorage, AK, USA, 1998, pp. 1932-1936 vol.3, doi: 10.1109/IJCNN.1998.687154.

12. G. S. Nagaraja, A. B. Soppimath, T. Soumya and A. Abhinith, "IoT Based Smart Agriculture Management System," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), Bengaluru, India, 2019, pp. 1-5, doi: 10.1109/CSITSS47250.2019.9031025.

13. A. K. Mariappan and J. A. Ben Das, "A paradigm for rice yield prediction in Tamilnadu," 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), Chennai, India, 2017, pp. 18-21, doi: 10.1109/TIAR.2017.8273679.

14, J. Zong and Q. Zhu, "Apply Grey Prediction in the Agriculture Production Price," 2012 Fourth International Conference on Multimedia Information Networking and Security, Nanjing, China, 2012, pp. 396-399, doi: 10.1109/MINES.2012.78.

15. X. Dai and H. Huang, "Modeling of Biology Catastrophe Prediction in Intelligent Agriculture," Second Workshop on Digital Media and its Application in Museum & Heritages (DMAMH 2007), Chongqing, China, 2007, pp. 253-257, doi: 10.1109/DMAMH.2007.9.

16. R. Medar, V. S. Rajpurohit and S. Shweta, "Crop Yield Prediction using Machine Learning Techniques," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033611.

17. J. Treboux and D. Genoud, "High Precision Agriculture: An Application Of Improved Machine-Learning Algorithms," 2019 6th Swiss Conference on Data Science (SDS), Bern, Switzerland, 2019, pp. 103-108, doi: 10.1109/SDS.2019.00007.

18. Y. Gandge and Sandhya, "A study on various data mining techniques for crop yield prediction," 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysuru, India, 2017, pp. 420-423, doi: 10.1109/ICEECCOT.2017.8284541.