International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

CROP YIELD PREDICTION USING MACHINE LEARNING

D. Joseph Pushparaj
Assistant Professor . Department of Computer Science and Engineering
PSN College of Engineering and Technology, Tirunelvei, Tamil Nadu, India

S. Ponnu Lakshmi
UG Student. Department of Computer Science and Engineering
PSN College of Engineering and Technology, Tirunelvei, Tamil Nadu, India

***Abstract*-**India is an Agriculture based economy whose most of the GDP comes from farming. The motivation of this project comes from the increasing suicide rates in farmers which may be due to low harvest in crops. Climate and other environmental changes have become a major threat in the agriculture field. Machine learning is an essential approach for achieving practical and effective solutions for this problem. Predicting yield of the crop from historical available data like weather, soil, rainfall parameters and historic crop yield. We achieved this using the machine learning algorithm. We did a comparative study of various machine learning algorithms, i.e., ANN, K Nearest Neighbour, Random Forest, SVM and Linear Regression and chose Random Forest Algorithm which gave an accuracy of 95%. In this project a web application has been developed which predicts the crop yield in general and also for a particular crop. Along with that, it also suggests the user if it is the right time to use the fertilizer or not.

Index Terms- ANN, SVM, Random Forest

I. INTRODUCTION

Across the globe India is the second largest country having a population of more than 1.3 Billion. Many people are dependent on agriculture but the sector lacks efficiency and technology especially in our country. By bridging the gap between traditional agriculture and data science, effective crop cultivation can be achieved. In developing countries, farming is considered as the major source of revenue for many people. In modern years, agricultural growth is engaged by several innovations, environments, techniques and civilizations. In addition, the utilization of information technology may change the condition of decision making and thus farmers may yield the best way. For the decision making process, data mining techniques related to agriculture are used. Data Mining is the process of analyzing, extracting and predicting the meaningful information from huge data to extract some pattern. This process is used by companies to turn the raw data of their customer to useful information. The process of Data Mining includes first selection of data followed by pre- processing of data and then transforming the data to get patterns which can then be used to predict useful insights. Preprocessing includes finding outliers and detecting missing values whereas transformation finds the correlation between objects. Applying the data mining techniques on historical climate and crop production data several predictions can be made on the basis of knowledge gathered which in turn can help in increasing crop productivity.

The current climatic and rainfall received data can be fetched. All those values get uploaded towards R. The captured data get analyzed with the agriculture dataset which helps to predict the climatic changes from the previous year. This will give a solution of cultivating which crop will be more profitable for the farmer. The changes in rainfall, humidity, temperature and pH level can be detected by using KNN, SVM and random forest. Data mining tools predict upcoming trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The motorized, prospective analyses offered by data mining move beyond the analyses of past events provided by display tools typical of decision support systems. Data mining tools can answer business questions that usually were too time-consuming to resolve. As per the rainfall and temperature range the district get spitted. The algorithm begins with the original set as the root node (rainfall). The decision tree is constructed with each non-terminal node (internal node - temperature range) representing the selected attribute on which the data was split, and terminal nodes (leaf nodes - district) representing the class label of the final subset of this branch.

II. Literature Survey

[1] K.K Joshi and P. Patidar, "An Approach to Construct Decision Tree using Sliq and Knn for Land Grading System", ISSN: 2277-5528, Impact Factor: 2.745 (Sijf), pp: 231-236.

Joshi K. K., and Patidar P., presented a method on Agricultural Data mining. Feature selection techniques were applied and relevant attribute only selected for further use. Some feature selection techniques were used ID3 Decision Tree Basics, Entropy, Information Gain and Supervised Learning in Quest (Sliq) Algorithm. They compared those feature selection techniques and results are visualized.

[2] Brijain R Patel and Kushik K Rana, "A Survey on Decision Tree Algorithm for Classification" The International journal of Engineering development and research, Vol. 2, Issue 1, 2014

Brijain R Patel and Kushik K Rana explained on classification and prediction are the techniques used to make out important data classes and predict probable trend .The Decision Tree is an important classification method in data mining classification. It is commonly used in marketing, surveillance, fraud detection, scientific discovery.

[3] Prachi H. Kulkarni and pratik D. Kute, " Internet Of Things Based System For Remote Monitoring of Weather Parameters and Applications", International Journal of Advances in Electronics and Computer Science, Vol. 3, Issue 2, February 2016

Prachi H. Kulkarni shared about the proof of concept for an IoT device that collects data regarding physical parameters, using a sophisticated microcontroller platform, from various types of sensors, through different modes of communication and then uploads the data to the Internet. The presented device has been designed for remote monitoring of weather parameters.

III. EXISTING SYSTEM

In existing system the Blind persons depend on others to move Farmers are struggling to produce the yield because of unpredictable climatic changes, decrease / increase in rainfall and drastically decrease in water supply. So that an agricultural data has been collected, stored and analysed for useful information. It is used to promote new advanced methods and approaches such as data mining that can give the information of the previous results to the crop yield estimation.

The analysis of a huge set of agriculture data is the major challenge in the agriculture and analysed for useful information classify the agriculture crops based on temperature changes and rainfall received range. The relevant features are selected and the crops are classified based on location. In this research work, the crop yield is estimated, and the most excellent crop can be chosen by analysing the climatic changes with the previous year data. The value and gain of the farming area can be improved using classification techniques.

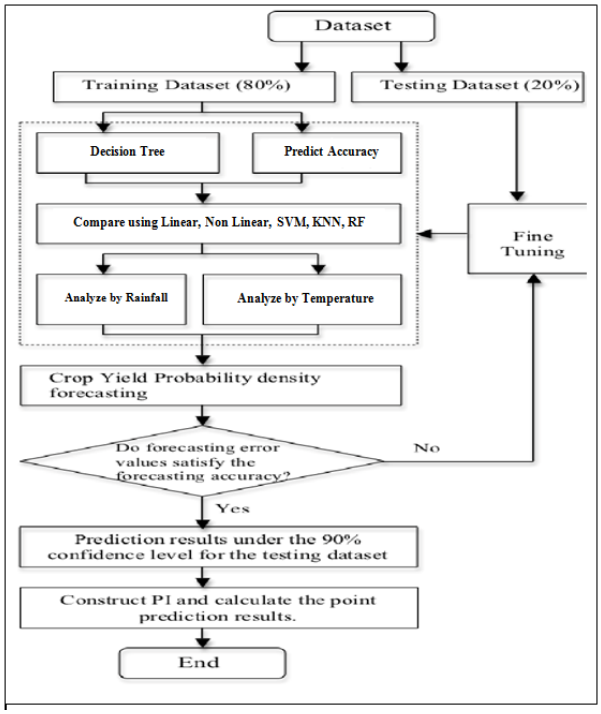
Disadvantage of Existing System

- The yield of crop can be reduced due to temperature and climate changes which can't be predict by the farmer.
- The farmer won't get good profit in crop yielding.

IV.PROPOSED SYSTEM

- As per the changes in temperature and rainfall, the crop can be predicted for cultivation which gives better profit for farmer.
- The changes in rainfall, humidity, temperature and pH level can be detected by using KNN, SVM and random forest. Those algorithms are compared to get better result in predicting the crop.
- The dataset contains all data related to past crop and climate changes, those data can be applied to analyze better crop selection.

V.IDENTIFY, PROJECTAND COLLECT IDEA



The current climatic and rainfall received data can be fetched. All those values get uploaded towards R. The captured data get analyzed with the agriculture dataset which helps to predict the climatic changes from the previous year. This will give a solution of cultivating which crop will be more profitable for the farmer. The changes in rainfall, humidity, temperature and pH level can be detected by using KNN, SVM and random forest. As per the changes in temperature and rainfall, the crop can be predicted for cultivation which gives better profit for farmer. The changes in rainfall, humidity, temperature and pH level can be detected by using KNN, SVM and random forest. Those algorithms are compared to get better result in predicting the crop. The dataset contains all data related to past crop and climate changes, those data can be applied to analyze better crop selection.

Dataset

- https://ourworldindata.org/crop-yields
- https://www.kaggle.com/datasets/thedevastator/the-relationship-between-crop-production-and-cli

This dataset provides data on crop yields, harvested areas, and production quantities for wheat, maize, rice, and soybeans. Crop yields are the harvested production per unit of harvested area for crop products. In most cases yield data are not recorded but are obtained by dividing the production data by the data on the area harvested. The actual yield that is captured on a farm depends on several factors such as the crop's genetic potential, the amount of sunlight, water, and nutrients absorbed by the crop, the presence of weeds and pests. This indicator is presented for wheat, maize, rice, and soybean. Crop production is measured in tonnes per hectare.

Module Description

Data Noise Removal

Data Preprocessing is a method that is used to convert the raw data into a clean data set. The data are gathered from different sources, it is collected in raw format which is not feasible for the analysis. By applying different techniques like replacing missing values and null values, we can transform data into an understandable format. The final step on data preprocessing is the splitting of training and testing data. The data usually tend to be split unequally because training the model usually requires as much data- points as possible. The training dataset is the initial dataset used to train ML algorithms to learn and produce right predictions (Here 80% of dataset is taken as training dataset).

Analyze Crop Affection

There are a lot of factors that affects the yield of any crop and its production. These are basically the features that help in predicting the production of any crop over the year. In this project we include factors like Temperature, Rainfall, Area, Humidity and Windspeed

Crop Yield Prediction

Before deciding on an algorithm to use, first we need to evaluate and compare, then choose the best one that fits this specific dataset. Machine Learning is the best technique which gives a better practical solution to crop yield problem. There are a lot of machine learning algorithms used for predicting the crop yield. In this project we include the following machine learning algorithms for selection and accuracy comparison :

Logistic Regression:-Logistic regression is a supervised learning classification algorithm used to predict the probability of target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. When logistic regression algorithm applied on our dataset it provides an accuracy of 87.8%.

Naive Bayes:-Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity,

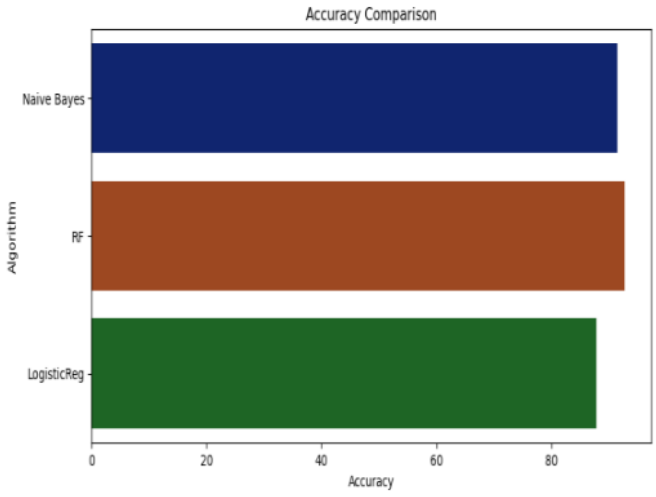
Naive Bayes is known to outperform even highly sophisticated classification methods. It provides an accuracy of 91.50%.

Random Forest:- Random Forest has the ability to analyze crop growth related to the current climatic conditions and biophysical change. Random forest algorithm creates decision trees on different data samples and then predict the data from each subset and then by voting gives better solution for the system. Random Forest uses the bagging method to train the data which increases the accuracy of the result. For our data, RF provides an accuracy of 92.81%

Classifier of Crops

Machine learning classifiers used for accuracy comparison and prediction were Logistic Regression, Random Forest and Naive Bayes. These three classifiers were trained on the dataset

and a comparison graph was plotted to showcase the performance of the models. Fig.5 showcase the performance of the models. Of the three classifiers used, Random Forest resulted in high accuracy.



Crop Name Prediction

Random Forest Classifier having the highest accuracy was used as the midway to predict the crop that can be grown on a selected district at the respective time. The preprocessed dataset was trained using Random Forest classifier. Chosen district's instant

weather data accessed from API was used for prediction. Trained model resulted in right crop prediction for the selected district.

ALGORITHM	ACCURACY
RANDOM FOREST	92.81407991690006
NAÏVE BAYES	91.49621790098573
LOGISTIC REGRESSION	87.82982929223341

Crop Yield Calculation

The crop which was predicted by the Random Forest Classifier was mapped to the production of predicted crop. Then the area entered by the user was divide from the production to get crop yield. Crop name predicted with their respective yield helps farmers to decide correct time to grow the right crop to yield maximum result.

$$Yield = Production/Area$$

VI. Working Methodology

CLASSIFICATION

The data mining function classification is used to assign items in a collection which helps to find target categories or classes. Classification places an important data mining task. The role of the classification is to predict the target class for each analyzing. The classification algorithm sets a main goal to increase the predictive accuracy obtained by the classification model.

Classification algorithms can follow three types of learning approaches in data mining: supervised learning, unsupervised learning and semi-supervised learning. The several classification techniques for discovering knowledge are

- **Rule Based Classifiers,**
- **Bayesian Networks (BN),**
- **Genetic Algorithms,**
- **Decision Tree (DT),**
- **Nearest Neighbor (NN),**
- **Artificial Neural Network (ANN),**
- **Support Vector Machine (SVM),**
- **Rough Sets,**
- **Fuzzy Logic,**
- **Random Forests.**

DECISION TREE

Decision Tree algorithm falls under the category of supervised learning algorithms. The supervised learning algorithm gives solution in an accurate prediction level, the decision tree algorithm can be used for predicting regression and classification **problems**. The prediction of crop as per the climate changes and temperature range can be done by using Decision Tree algorithm. The training model for predicting a crop can be constructed by using Decision Tree. The target variables for the class can be derived by **learning decision rules** concluded from prior data(training data). The training data contains field such as state, district, crop year, season, annual rainfall, minimum temperature, maximum temperature and crop. The tree model get constructed by using decision tree algorithm. Each internal node represents various attributes of the data set and the leaf node corresponds to a class label.

Decision Tree Algorithm processing step

1. Fix the best or initial prediction attribute of the dataset at the root of the tree.
2. Various subsets should split from the training set. The split subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until finding leaf nodes in all the branches of the constructed decision tree.

Decision Tree construction steps

- In the beginning level, the whole training set will be taken as the root.
- The interior values are referred to be categorical. If the values in the construction of tree are continuous then they are discretized prior in building the model.
- As per the attribute values, the records are distributed recursively.
- Using statistical approach, the attribute get placed as root and internal node.

The prediction of class label for a record will start from the root follow by internal nodes. The root attribute place a high priority in predicting the suitable crop. As per the need, the comparison will jump to the next internal node. By comparing temperature and rainfall attribute values with other **internal nodes** of the tree, the flow will reach a leaf node **(crop)** with predicted class value.

ATTRIBUTE SELECTION MEASURES

During model construction it will be a complicated step in deciding the attribute to be place at root and different levels of the tree. The random selection in placing a attribute won't solve the issue and prediction of crop as per the climatic changes can't

be accurate. In decision tree, a attribute for the root and internal node can be selected by any one measure as follows:

- **Information Gain**
- **Gini Index**

Information Gain

In this research the prediction of crop as per the climatic condition was carried out by using information gain measure. The information gain helps to estimate the information contained by each attribute. To measure the randomness or uncertainty of a random variable X is defined by Entropy.

$$H(X) = \mathbb{E}_X[I(x)] = -\sum_{x \in \mathbb{X}} p(x) \log p(x).$$

The information gain can be calculated entropy measure of each attribute. The attribute get sorted which helps in finding information gain at expected reduction in entropy.

VI. RESULT

Dataset -1

datainput - DataFrame

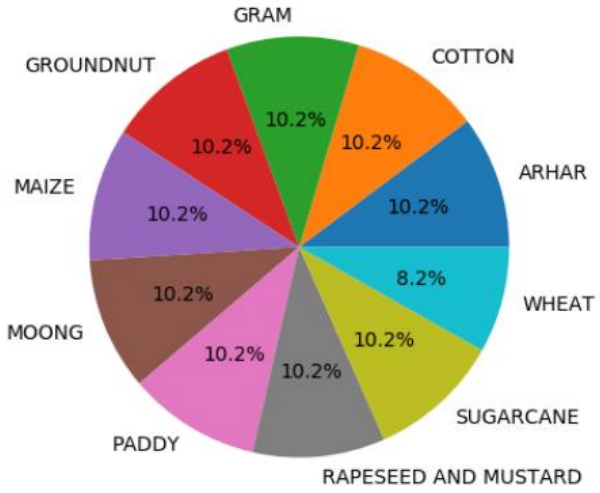
Index	Crop	State	ultivation (`/Hecta	Cultivation (`/Hec	Production (`/Qui	d (Quintal/ Hecta	Support price
0	ARHAR	Uttar Pradesh	9794.05	23076.7	1941.55	9.83	6000
1	ARHAR	Karnataka	10593.1	16528.7	2172.46	7.47	6000
2	ARHAR	Gujarat	13468.8	19551.9	1898.3	9.59	6000
3	ARHAR	Andhra Pradesh	17051.7	24171.7	3670.54	6.42	6000
4	ARHAR	Maharashtra	17130.5	25270.3	2775.8	8.72	6000
5	COTTON	Maharashtra	23711.4	33116.8	2539.47	12.69	5515
6	COTTON	Punjab	29047.1	50828.8	2003.76	24.39	5515
7	COTTON	Andhra Pradesh	29140.8	44756.7	2509.99	17.83	5515
8	COTTON	Gujarat	29616.1	42070.4	2179.26	19.05	5515
9	COTTON	Haryana	29919	44018.2	2127.35	19.9	5515

Dataset -2

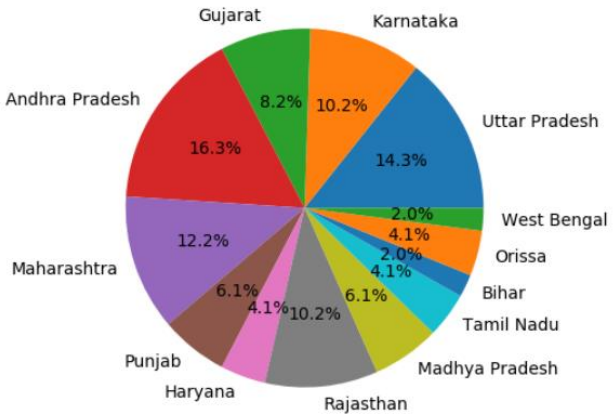
datainput - DataFrame

Index	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254	2000
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2	1
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102	321
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176	641
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720	165
5	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Coconut	18168	6.51e+07
6	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Dry ginger	36	100
7	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sugarcane	1	2
8	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Sweet potato	5	15
9	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Tapioca	40	169
10	Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Arecanut	1254	2061
11	Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Other Kharif pulses	2	1
12	Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Rice	83	300
13	Andaman and Nicobar Islands	NICOBARS	2001	Whole Year	Cashewnut	719	192

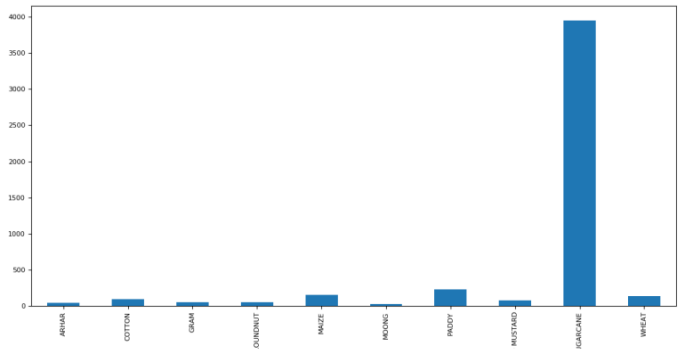
Crop-wise distribution in percentage



State wise distribution in dataset



Total Yield crop wise



Classification Report

Class	Precision	Recall	F1 Score	Support (num of examples)
0	0.78	1.00	0.88	7
1	1.00	0.75	0.86	8
Accuracy	0.87			15

Overall Result

Algorithm	Precision		Recall		F1 Score		Accuracy
	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1	
Logistic Regression	1.0	0.89	0.86	1.0	0.92	0.94	0.93
Decision Tree	0.78	1.0	1.0	0.75	0.88	0.86	0.87
Random forest	0.86	0.75	0.75	0.86	0.80	0.80	0.80
K nearest	0.50	0.86	0.75	0.67	0.60	0.75	0.69

Algorithm Result

Algorithm	R2 score	Mean absolute error
Decision Tree	0.84	167163.3086041714
Random Forest	0.91	155503.99436675265

VII. CONCLUSION

Data mining is the latest research area of agriculture. This is reasonably a fresh research field and it is projected to grow in the future. Nowadays, farmers are struggling to produce the yield because of unpredictable climatic changes and range of rainfall. The experiments conducted to analyze a small number of traits contained within the dataset to decide their effectiveness when compared with standard statistical techniques. Feature selection is very important to classify the crop data. In this research work, decision tree concept was used to build a model in finding a solution of which crop is better for cultivation. ID3 algorithm is used to classify the crops and a solution to increase their yield by analyzing the climatic changes. This work helps to predict the best crop for analyzing the temperature and rainfall. This will help farmers to increase the yield and income level. The obtained accuracy of ID3 algorithm is 95.83 % and it is better than other algorithms.

REFERENCES

[1] M. Alagurajan, and C. Vijayakumaran, "ML Methods for Crop Yield Prediction and Estimation: An Exploration," International Journal of Engineering and Advanced Technology, vol. 9 no. 3, 2020

[2] P. Kumari, S. Rathore, A. Kalamkar, and T. Kambale, "Predicition of Crop Yeild Using SVM Approch with the Facility of E-MART System" Easychair 2020. [6] S. D. Kumar, S. Esakkirajan, S. Bama, and B. Keerthiveena, "A microcontroller based machine vision approach for tomato grading and sorting using SVM classifier," Microprocessors and Microsystems, vol. 76, pp.103090, 2020

[3] P. Tiwari, and P. Shukla, "Crop yield prediction by modified convolutional neural network and geographical indexes," International Journal of Computer Sciences and Engineering, vol. 6, no. 8, pp. 503-513, 2018.

[4] P. Sivanandhini, and J. Prakash, "Crop Yield Prediction Analysis using Feed Forward and Recurrent Neural Network," International Journal of Innovative Science and Research Technology, vol. 5, no. 5, pp. 1092-1096, 2020.

[5] N. Nandhini, and J. G. Shankar, "Prediction of crop growth using machine learning based on seed," Ictact journal on soft computing, vol. 11, no. 01, 2020

[6] A. A. Alif, I. F. Shukanya, and T. N. Afee, "Crop prediction based on geographical and climatic data using machine learning and deep learning", Doctoral dissertation, BRAC University) 2018.

[7] A. Fuentes, S. Yoon, S. C. Kim, and D. S. Park, "A robust deeplearning-based detector for real-time tomato plant diseases and pests' recognition," Sensors, vol. 17, no. 9, pp. 2022, 2017.

[8] J. Sun, L, Di, Z. Sun, Y. Shen, and Z. Lai, "County-level soybean yield prediction using deep CNN-LSTM model," Sensors, vol. 19, no. 20, pp. 4363, 2019.

[9] K. A. Shastry, and H. A. Sanjay, "Hybrid prediction strategy to predict agricultural information," Applied Soft Computing, vol. 98, pp. 106811, 2021.

[10] D. A. Bondre, and S. Mahagaonkar, "Prediction of Crop Yield and Fertilizer Recommendation Using Machine Learning Algorithms," International Journal of Engineering Applied Sciences and Technology, vol. 4, no. 5, pp. 371-376, 2019.