IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 6, May 2023

Crop Yield Prediction using Machine Learning

Shubham Khade[1], Vishal Malaye[2], Saurabh Mhase[3], Ashutosh Shitole[4], Prof. Shraddha Shirsath[5]

Student, Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, India[1,2,3,4]

Professor, Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune, India[5]

Abstract: *In India, agriculture is the main occupation. It the primary source of income of the population. 60% of the population is dependent on agriculture related to employment. Thereby, agriculture is considered as the backbone of the country and helps in economic growth. Agriculture is a primary sector occupation and provides raw materials to secondary sector which includes food factories, textile industry, food security, etc. So, agricultural productivity is largely dependent to increase the economy of the country, leading to the business growth and provide employment to most of the population. Productivity largely depends on climatic factors and environmental conditions such as rainfall, temperature, humidity, soil type, etc. Due to unfavourability, the crop yield production gets affected thereby harming the economy. Using appropriate machine learning models helps to predict the crop yield considering climatic conditions. This paper is based on three algorithms viz. polynomial regression i.e., linear regression, support vector regression and random forest regression. These algorithms will help in suitable crop selection to grow.The selecting of crop is very important because it will provide us the most productivity hence increasing profit. This research work will help the farmers and producers increase the productivity and boost the economy.*

Keywords: *Agriculture, Economy, Productivity, Machine Learning Algorithms, Predict Crop Yield.*

I. INTRODUCTION

Agriculture is one of the most important sectors for human survival and economic growth. Use of crop yield prediction system is the most important task for farmers, policymakers, and other stakeholders in the agriculture industry. Accurate prediction of crop yield can help farmers plan their planting schedules, optimize the use of resources such as water, fertilizer, and pesticides, and make informed decisions about crop management practices. In recent years, the use of machine learning techniques for crop yield prediction has gained significant attention. Machine learning models can be trained to analyze large datasets and identify patterns and relationships between various factors that affect crop yield. With the availability of high-quality data on weather conditions, soil quality, and crop types, machine learning algorithms can provide accurate predictions of crop yields.

In this project, we aim to develop a machine learning model for crop yield prediction based on various environmental factors. We will use historical data on crop yields, weather patterns, and soil characteristics to train and test the model.We will evaluate different machine learning algorithms and identify the most accurate and reliable model. The final model will be used to predict crop yield for a given set of environmental conditions and can assist farmers in making informed decisions about crop management and resource allocation.

II. MOTIVATION

The motivation behind the proposed system is to leverage technology, specifically machine learning, to optimize crop selection and increase agricultural productivity in India. Traditional farming methods are facing challenges due to climate change and unfavourable weather conditions, resulting in significant losses in crop yield. By using machine learning models to predict unfavourable conditions and provide data-driven recommendations on crop selection, irrigation methods, and other best practices, farmers can make informed decisions and mitigate losses. The goal is to increase crop yield, boost the economy of the country, and improve the livelihood of farmers.

III. OVERVIEW

India is primarily an agrarian country, with agriculture serving as a crucial source of livelihood and economic growth. However, climate change, unfavourable weather conditions, and other factors can lead to loss of crop yield. The

DOI: 10.48175/IJARSCT-10102

proposed system aims to address these challenges by utilizing machine learning algorithms, such as polynomial regression, random forest, and support vector machine regression, to create predictive models. These models will consider various factors, including climatic conditions, soil type, irrigation methods, and geographical factors, to provide farmers with recommendations on which crops to grow in a particular region. This will help farmers optimize crop selection, increase productivity, and reduce losses caused by unfavourable weather conditions, ultimately benefiting the agricultural sector and the overall economy of the country.

IV. LITERATURE SURVEY

In [1] J.P. Singh, Rakesh Kumar, M. P. Singh, and Prabhat Kumar, have concluded that the research work done by them helps in perfectly improving the yield rate of crops by applying some classification methods and comparing the parameters. We can also do analyzing and prediction of crops using Bayesian algorithms. The algorithms used in their research work are Bayesian algorithm, K-means Algorithm, Clustering Algorithm, Support Vector Machine. The disadvantage is that there is no proper accuracy and performance.

P.Priya et al. [2], focuses more on supervised learning methods like random forest classifier and decision tree classifier. Here, the datasets considered consists of rainfall, perception, production, temperature, and a number of decision trees by including more than half of the records in the datasets. The application of decision trees on the remaining records takes place for increase in accuracy rate after classification. From this paper, we came to know that assembling of decision tree with RF may give better results with more accuracy. Limitation: Can use more complex algorithms for assembling.

Arun Kumar et al. [3] highlights the weightage of Support Vector machine algorithm in the crop yield estimation. Crop yield is performed to categorize based on yield productivity and class labels. The parameters considered are, variation of crop yield with rainfall, variation of humidity factor and the impact of climatic change on agriculture. ARIMA model is also used to operate on data with time series. Limitation: Less parameters are taken into consideration and Time series Analysis may give less accuracy than ensemble techniques.

In [4] the authors Subhadra Mishra, Debahuti Mishra and Gour Hari Santra, have concluded that this is an advanced researched field and is expected to grow in the future. The integration of computer science with agriculture helps in forecasting agricultural crops. This method also helps in providing information of crops and how to increase yield rate. The algorithms used are Artificial neural networks, Decision Tree Algorithms, Regression analysis. The disadvantage is clear methodology is not specified

In [5] Nishit Jain, Amit Kumar, Sahil Garud, Vishal Pradhan, Prajakta Kulkarni, have concluded in their research work that their paperwork helps in predicting crop sequences and maximizing crop profit yield rates and as a result, making benefits to the farmers. Also, using machine learning algorithm applications with agriculture in predicting crop diseases, studying crops using simulations, different irrigation patterns. The algorithms used are Artificial neural networks, Support Vector Machine. The disadvantage is that the precision accuracy is not specified.

In [6] B.Mallikarjun Rao, D.Sindhura, B.Navya Krishna, K.Sai Prasanna Lakshmi, Dr. J Rajendra Prasad, have concluded that this method will provide a useful and accurate knowledge. Using this knowledge, we predict and support the decision making for different sectors. The algorithms used are multiple linear regressions. The disadvantage is that it can be applied for limited areas.

Balamurugan [7], in his research work have implemented crop yield prediction system by using only one algorithm, the random forest classifier. Various features like rainfall, temperature and seasons were taken into account to predict the crop yield. Other machine learning algorithms were not used and applied to the datasets. With the absence of other algorithms, comparison and quantification were missing in the work thus, unable to provide the better prediction system.

Chawla, I. et al. [8] (2019, August) used fuzzy logic for crop yield prediction through statistical time series models. The research work considered parameters like rainfall and temperature for prediction. Their prediction was classification with levels ‘good yield’, ‘very good yield’

Kalbande, D. R. et al. [9] (2018) used support vector regression, multi polynomial regression and random forest regression for prediction of corn yield and evaluated the model’s using metrics like errors namely MAE, RMSE and R-square values

Singh, C. D. et al. [10]. (2014, January) developed an application to advise crops which works on selected districts of Madhya Pradesh, India. The user has to provide input about cloud cover, rainfall, temperature, observed yield in the past crop growth and the system would predict the yield. Depending on the trigger values set that has been obtained, the crop will be labeled some particular value and obtain the results.

Dr. Y. Jeevan Nagendra Kumar [11], have concluded in his research work that Machine Learning algorithms can predict a target/outcome about crops by using Supervised Learning algorithms. The research work focuses on supervised learning techniques for crop yield prediction. To get the specified expected resulting outputs, it needs to generate an appropriate function by some set of variables which can map the input variable to the aim output. The research work consists of using Random Forest algorithm and the paper conveys that by using this algorithm for prediction, the best accurate values are obtained by considering least number of models.

In [12].Machine learning approach for predicting crop yield based on parameters of climate conditions. The paper has been provided in International Conference on Computer Communication and Informatics (ICCCI). In the current research a software tool named Crop Advisor has been developed as a user-friendly web page for predicting the influence ofclimatic parameters on the crop yields. C4.5 algorithm is used to predict the crop yields of selected crops in selected districts of Madhya Pradesh and produce the most influencing climatic parameter on the crop yield. The research work carried out and the paper is implemented using Decision Tree.

In[13].RandomForestsmachine learning algorithm is used forGlobalandRegionalCropYieldPredictions from the institute on the Environment, University of Minnesota, St.Paul, MN 55108, United States of America. The output generated shows that Random Forest is an effective machine learning algorithm for crop yield predictions at regional and global scales for its highaccuracy.Thepaperhas beenimplementedusingk-nearest neighbor, Support Vector Regression and Random Forest.

[14]. "Crop Yield Prediction Based on Deep Learning: A Case Study of Maize Yield Prediction in China" (2021) by Y. Wang et al.: The authors carried out the research work using deep learning approach based on convolutional neural networks (CNNs) to predict maize yield in China. They used remote sensing data, meteorological data, and soil data as input features to the model. Their approach achieved an accuracy of 92.8% in predicting maize yield.

[15]."Agricultural Crop Yield Prediction Using Machine Learning Techniques: A Review" (2019) by V. Singh and S. K. Sharma: The authors reviewed several machine learning techniques for crop yield prediction, including linear regression, support vector regression, decision trees, and random forests. They compared the performance of these techniques on different crop datasets and highlighted the advantages and disadvantages of each approach.

[16]. "Crop Yield Prediction Using Machine Learning: A Review" (2019) by M. M. Z. Hossain et al.: The authors reviewed recent works on crop yield prediction using machine learning, focusing on different input data sources such as remote sensing data, meteorological data, and soil data. They also discussed the challenges and limitations of these approaches, such as data quality and availability issues.

[17]. "Deep Learning for Crop Yield Prediction Based on Remote Sensing Data" (2018) by Y. Yao et al.: The authors proposed a deep learning approach based on CNNs to predict crop yield using remote sensing data. They used the

DOI: 10.48175/IJARSCT-10102

normalized difference vegetation index (NDVI) as input features to the model. Their approach achieved an accuracy of 94.6% in predicting rice yield.

[18]. "Crop Yield Prediction Using Random Forest Regression: A Case Study of Maize Yield Prediction in China" (2020) by Y. Wang et al.: The authors used a random forest regression approach to predict maize yield in China. They used remote sensing data, meteorological data, and soil data as input features to the model. Their approach achieved an accuracy of 92.6% in predicting maize yield.

V. ALGORITHM

Linear Regression:

Linear Regression is one of the machine learning algorithms which based on supervised learning method and regression task is performed by it. Regression model helps to get a target prediction value which is based on independent variables. Mostly, it is used to find the relationship between variables and forecasting. Different regression models differ based onthe kind of relationship between independent and dependent variables and the number of independent variables being used.

The task to predict a dependent variable (y) based on a given independent variable (x) is performed by linear regression. So, this regression technique is used to find a linear relationship between x (input) and y(output). Therefore, the name is Linear Regression.

Hypothesis function for Linear Regression:$y = \theta 1 + \theta 2$

Working of Linear Regression Algorithm

Linear Regression is a machine learning algorithm used for predicting a continuous target variable based on one or more predictor variables. It works by finding a linear relationship between the predictor variables and the target variable. Here's how it works:

1. **Data Pre-processing:** The data is pre-processed by scaling and normalizing the features to ensure that they have similar ranges.

2. **Model Training:** A linear equation of the form $y = mx + b$ is fitted to the training data, where y is the target variable, x is the predictor variable, m is the slope or coefficient, and b is the intercept. The model tries to find the values of m and b that best fit the data.

3. **Cost Function Optimization:** A cost function is defined to measure the difference between the predicted values and the actual values. The goal is to minimize the cost function by adjusting the values of m and b. The most common cost function used in linear regression is the mean squared error (MSE).

4. **Model Evaluation:** The model is evaluated on a separate test dataset to measure its performance. Common metrics used to evaluate a linear regression model are the coefficient of determination (R^2), mean squared error (MSE), and root mean squared error (RMSE).

5. **Prediction:** Once the model is trained, it can be used to make predictions on new data by applying the linear equation to the predictor variables.

Linear Regression has several advantages over other regression algorithms:

- It is simple.
- Easy to implement.
- It provides interpretable results, as the coefficients of the predictor variables can be used to understand the relationship between the variables and the target variable.
- It can handle both categorical as well as continuous predictor variables.
- It can be used for both simple and complex regression tasks.

Support Vector Regression:

Support Vector Regression is a type of supervised learning algorithm. It is used to predict discrete values. Support Vector Regression uses the same principle as Support Vector Machine(SVM). The basic idea behind Support Vector Regression(SVR) is to find the best fit line. The best fit line in the Support Vector Regression is the hyperplane which has maximum number of points. The other Regression models try to minimize the error between the real and predicted values, while the SVR tries to find the best fit line within a threshold value. The distance between the hyperplane and boundary line called as a threshold value. The best time complexity of SVR which should be the best fit line, which is more than the quadratic with the number of samples that makes it hard to scale for the datasets, with more than couple of 10000 of samples.

Working of Support Vector Regression Algorithm

Support Vector Regression (SVR) is a machine learning algorithm that is used for regression tasks. It is based on Support Vector Machines (SVMs) and uses the same principles of maximizing the margin while finding a hyperplane that separates the data into different classes. However, in SVR, the goal is to find a hyperplane that best fits the data instead of separating it.

Here's how SVR works:

1. **Data Pre-processing:** The data is pre-processed by scaling and normalizing the features to ensure that they have similar ranges.

2. **Kernel Selection:** A kernel function is selected to transform the input data into a higher-dimensional feature space. This transformation allows the data to be separated by a hyperplane that is nonlinear in the original input space. Common kernel functions used in SVR are linear, polynomial, radial basis function (RBF), and sigmoid.

3. **Parameter Tuning:** Two important parameters in SVR are C and ε. C controls the trade-off between the complexity of the model and the amount of error that is allowed in the training set. A smaller value of C will result in a simpler model with more error, while a larger value of C will result in a more complex model with less error. ε controls the width of the epsilon-insensitive zone. A smaller value of ε will result in a tighter fitting of the data, while a larger value of ε will result in a looser fitting of the data.

4. **Model Fitting:** The SVR model is fitted to the training data by finding the hyperplane that best fits the data while satisfying the constraints of the margin and the epsilon-insensitive zone. The goal is to minimize the sum of the error between the predicted values and the actual values while keeping the margin as large as possible.

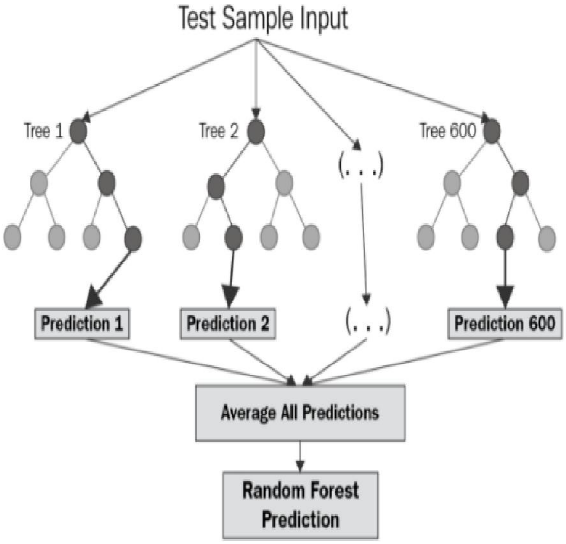
5. **Prediction:** Once the SVR model is trained, it can be used to make predictions on new data by applying the kernel function to transform the input data into the higher-dimensional feature space, and then using the hyperplane to predict the output values.

SVR has several advantages over other regression algorithms:

- It can handle nonlinear relationships between features and target variables.
- It is less affected by outliers and noise in the data.
- It provides a sparse solution, which means that only a subset of the training data is used to build the model.
- It can handle high-dimensional data.
- It can handle large datasets.

Random Forest

Random Forest Regression is one of the supervised machine learning algorithms which makes use of ensemble learning method for regression. The predictions made by different machine learning algorithms are combined to make a more accurate prediction rather than using a single model. This technique is called ensemble learning method.



The structure of a Random Forest is shown in the above diagram. There is no interaction among the trees as they run in parallel. During training time, the Random Forest operates by constructing several decision trees and outputs the mean of the prediction of all the trees. To know more about Random Forest algorithm, lets walk through the steps:

- **a.** From the training set, pick a k random data point.
- **b.** By associating to k data points, build a decision tree.
- **c.** In order to build the number of trees, choose the number N and repeat steps a and b.
- **d.** For the new data point, make each one of your N-tree trees predict the value of y for data point in problem and provide the new data point for an average across the all the predicted y values. A Random Forest Regression model is accurate and powerful. It performs excellent role in solving many problems, including features with non-linear relationships. Limitations of this algorithm:interpret-ability is absent, over fitting occurs easily.

Working of Random Forest Algorithm

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy and stability of predictions. Here is how it works:

1. **Data Sampling:** A random subset of the training data is selected for each tree in the forest. This process is called bootstrapping. It helps to create diversity among the trees and prevent overfitting.

2. **Feature Selection:** At each node in a decision tree, a random subset of features is considered for splitting. This process is called feature bagging. It helps to reduce the correlation between trees and improve their independence.

3. **Tree Construction:** A decision tree is constructed using the selected features and the bootstrapped data. The tree is grown until a stopping criterion is met, such as maximum depth or minimum number of samples per leaf.

4. **Ensemble Aggregation:** The predictions of all trees in the forest are combined to produce a final prediction. The most common method of aggregation is to take the majority vote of the predictions. This method is called hard voting. Soft voting can also be used, where the probabilities of each class are averaged over all trees.

Random Forest Pseudocode:

1. Randomly select “k” features from total “m” features.Where k << m
2. Among the “k” features, calculate the node “d” using the best split point.

DOI: 10.48175/IJARSCT-10102

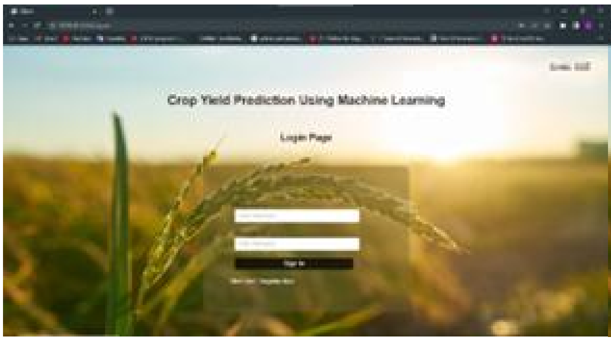
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until "l" number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.

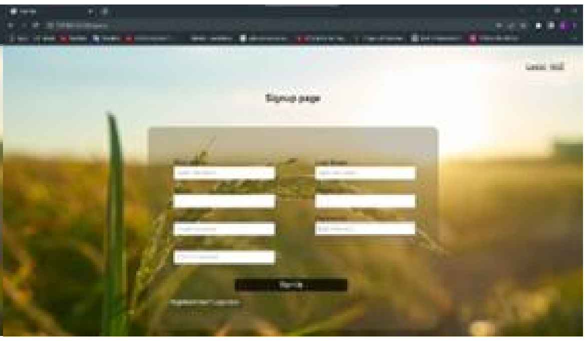
Random Forest has several advantages over a single decision tree:

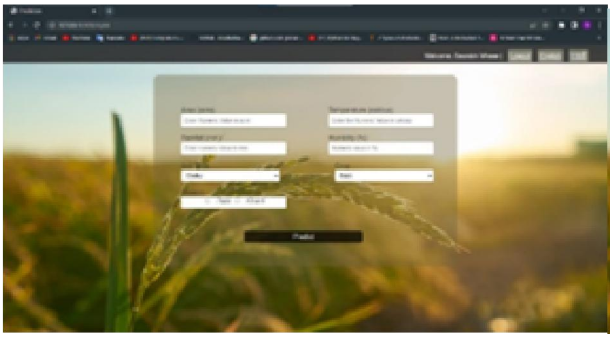
1. It is more accurate and stable, as it reduces overfitting and variance.
2. It can handle high-dimensional data and nonlinear relationships between features and target variables.
3. It can provide feature importance measures, which can be used for feature selection and interpretation.
4. It is easy to implement and parallelize, as each tree can be trained independently.

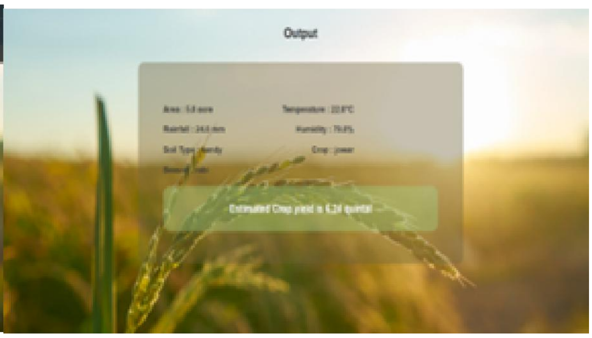
VI.SCREENSHOTS

After running our system, some of the results that we got are provided below.









Area : 3.0 acre
Temperature : 22.8°C
Rainfall : 24.0 mm
Humidity : 75.0%
Soil Type : Sandy
Crop : jowar
Season : Rabi
Estimated Crop yield is 6.24 quintal

VII. FUTURE WORK

1. Incorporating more data sources: While existing studies have demonstrated the effectiveness of machine learning models in predicting crop yield, there is a need for incorporating more data sources such as remote sensing and satellite data, to improve the accuracy of the models.
2. Transfer learning: Transfer learning is a machine learning technique that involves transferring knowledge from one model to another. In the context of crop yield prediction, this could involve transferring knowledge from models developed in one region to another region, where data is scarce.
3. Integrating models with precision agriculture technologies: Precision agriculture technologies such as drones, sensors, and GPS can provide real-time data on environmental conditions and crop performance, which can be used to improve the accuracy of crop yield prediction models.
4. Implementing models in decision support systems: Decision support systems can provide farmers with real-time recommendations on crop management practices such as irrigation, fertilization, and pesticide application.

DOI: 10.48175/IJARSCT-10102

Integrating crop yield prediction models into these systems can help farmers make informed decisions about crop management.

5. Predicting other crop metrics: While crop yield is an important metric, other metrics such as crop quality and disease resistance are also important for farmers. Developing machine learning models to predict these metrics can help farmers make more informed decisions about crop management.

VIII. CONCLUSION

Agriculture is the primary field of occupation in India, which helps in economic growth of our country. But this occupation is lacking behind because new technologies are not used. So, by using new technologies of machine learning we can have huge impact on economy. Hence our farmers should be aware of new technologies and get more benefit by becoming digitalized. Ultimately, by using machine learning algorithms, we have implemented a system which collects data and predict the result. The algorithms we have used are Support Vector Regression, Polynomial regression, and Random Forest Algorithm to predict the outcome with increased accuracy rate. Hence, in order to improve the performance and to get maximum yield, we have to compare the accuracy and prediction of different crops; and the crop which provides maximum yield will be grown for greater profit. Different crops give different results as per the prediction and by comparing the results of different crops, which helps the farmers to choose the best accuracy crop to grow for maximum yield.

REFERENCES

[1] Ms Kavita, Pratistha Mathur, "Crop Yield Estimation in India Using MachineLearning", 2020 IEEE 5thInternational Conference on Computing Communication and Automation (ICCCA) Galgotias University, Greater Noida, UP, India. Oct 30-31, 2020.

[2] Mummaleti Keerthana, K J M Meghana, Siginamsetty Pravallika, Dr. Modepalli Kavitha, "An Ensemble Algorithm for Crop Yield Prediction", Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021). IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-1183-4.

[3] Ramesh Medar, Vijay S. Rajpurohit, Shweta, "Crop Yield Prediction using Machine Learning Techniques" 2019 5thInternational Conference for Convergence in Technology (I2CT) Pune, India. Mar 29-31, 2019

[4] M.Kalimuthu, P.Vaishnavi, M.Kishore, "Crop Prediction using Machine Learning" Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020) IEEE Xplore Part Number: CFP20P17-ART; ISBN: 978-1-7281-5821-1

[5] Potnuru Sai Nishant, Pinapa Sai Venkat, Bollu Lakshmi Avinash, B. Jabber, "Crop Yield Prediction based on Indian Agriculture using Machine Learning" 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020.

[6] Aruvansh Nigam, Saksham Garg, Archit Agrawal, Parul Agrawal, "Crop Yield Prediction Using Machine Learning Algorithms" 2019 Fifth International Conference on Image Information Processing (ICIIP).

[7] D.Jayanarayana Reddy, Dr M. Rudra Kumar, "Crop Yield Prediction using Machine Learning Algorithm" Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021) IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2.

[8] Anakha Venugopal, Aparna S, Jinsu Mani, Rima Mathew, Prof. Vinu Williams," Crop Yield Prediction using Machine Learning Algorithms" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181.

DOI: 10.48175/IJARSCT-10102