**A Thorough analysis of Crop Prediction by using Random forest and SVM algorithms .**

Author :Atharva Karale,Pratik Mahajan,Shreyash Borade,Piyush Khunte

**Abstract**

This paper discusses the use of different machine learning techniques, particularly Random forest and SVM algorithms in crop forecasting, with regard to improving the decision-making processes undertaken by the farmers. The three models-in-use: Random Forest, SVM, and Logistic Regression are compared on a dataset that features soil nutrients such as nitrogen, phosphorus, potassium with environmental factors: temperature, humidity, pH., Rain). Crop type comes in 22 different types and in these tests, accuracy, precision, recall, and F1 score were analyzed for the models. Using hyperparameter tuning, Random Forest outperformed other models in terms of accuracy with 99.31%, followed by SVM with 97.73%, and logistic regression with 96.36%. The strongest forest model with the fine tremor can be found. Such findings justify the application of machine learning-based solutions in agriculture to enhance production forecasting and the decision that makes yield production. 1. Introduction

**1.1 Background**

Agribusiness has been at the helm throughout the era of human history. Even today, in the sophisticated world, it is perceived as an inevitable constituent of the world's economy, including developing countries.

The demand for food production increases as the population grows as well as the growing industrial sector. However, crop yields are often at a risk due to environmental vagaries including climatic change, soil deterioration, and lack of nutrient management. It is therefore important for farmers to have proper and reliable systems for deciding on what crops to plant under which conditions of soil and climate. The large amount of data analysis that has made machine learning highly powerful in discovering patterns and achieving very high accuracy in prediction, particularly in industries such as agriculture, came very handy.

Crop prediction happens to be one of the highly important applications of ML in agricultural endeavors. It guides farmers on the right type of crop to plant under specific environmental and soil conditions hence optimizing yields and reducing losses. During this research, we intend to compare the three machine learning algorithms; Random Forest, SVM, and Logistic Regression towards the prediction of the best crops for cultivation under key factors of the soil nutrient content and the weather condition.

**1.2. Research problem**

Despite significant technological breakthroughs in agriculture, one of the big challenges still left is the accurate prediction of crop yields in various regions due to different environmental factors.

The systems existing are based on traditional methods that do not consider the complexity of immediate or historical data with regard to soil composition and climate change. This builds an imbalance between the need for farmers and the solutions available. However, still important comparative studies are needed to determine which algorithm performs better in predicting the yield. It should be noted that previous related work has proved that RF yields a better result compared with the KNN simpler model; hence, more complex algorithms like SVM, though used, are currently employed without full-scale investigation for the agricultural datasets. This research seeks to bridge the gap by conducting an in-depth comparison between random forest, support vector machines, and logistic regression to assess their performance in yield prediction.

**1.3. Objectives**

The specific objectives of this paper are to:

Evaluate the performance of random forests, support vector machines, and logistic regression in predicting crop yields using soil and environmental factors.

Use GridSearchCV to fine-tune a random forest model to obtain optimal hyperparameter settings and thereby improve its prediction accuracy.

Compare the contribution of different characteristics, such as nitrogen content, moisture and precipitation, to yield predictions and assess their importance in the final model.

**2. Related works**

The application of machine learning to yield forecasting is not a novel concept, but with rapid development in computer technology, more sophisticated methods are now adopted.

Different studies have studied the use of algorithms such as decision-making in trees, knees, and random forests to predict the harvest yield. A comparative analysis on yield prediction was carried out by Rao et al. (2022) regarding Random Forest and KNN, which stated that the highest accuracy was obtained from the Random Forest model. Their paper has important suggestions regarding such traits like nitrogen, potassium, and moisture while determining yield. Kanaga Priya, etc. With MLP perception of multiple layers, they proposed a gradient separation method with the aid of IoT for providing the actual yield recommendations and thus improving real-time data collection on the field. For that, they reported getting better accuracy while integrating their real-time data into machine learning models. However, though these models have been proved to be very efficient, they often do not generalize well across other climatic zones and soil conditions and require a more comprehensive approach using robust algorithms such as SVM and Random Forest. Another study of Kalimuthu et al. As compared with naive Bayese, decision -making trees and random forests, it has been found that the performance of random forests is better than the other models in handing large large data sets. This study is based on these discoveries and aims at a much more deeply comparison between forest, SVM and logistical regression to assess their effectiveness in the harvest. Method

**3. Data set description**

This data set has 2200 observations and is described on 7 key variables:

N (nitrogen content): very essential to plant growth. P (phosphorus content): useful in the development of roots.

K (potassium content): provides plants with resistance to diseases. Temperature: measured in degrees Celsius.

Humidity: Relative humidity as percentage. pH: This measure shows the degree of soil acidity or alkalinity.

Precipitation: measured in millimeters. Besides that dataset, 22 other cultural labels are reported that match crops which can be planted based on the value of the function. The data is gathered from various agricultural countries to ensure the variety of soil and climatic conditions. Hence, it is crucial that the dataset does not contain any missing values to provide easy preprocessing.

**3.2. Machine Learning Models**

For this study, we are using three machine learning models:

**Random Forest**: It is ensemble learning based on the idea of using multiple decision trees in a form of an ensemble procedure. In other words, creating a large number of trees during training and making predictions based upon the majority of all trees. Random forests are particularly effective for high-dimensional data sets where the risk of overfitting can be reduced by averaging multiple predictions. Support Vector Machine **(SVM)**: SVM is a pretty robust algorithm that works on the idea of finding the hyperplan most suited for separating the data points in different classes. In this study, we have made use of a radial basic function (RBF) core that is very effective when the relationship between the functions is non-linear.

**Logistic Regression**: A logistic regression model is one of the most widely used linear models for classification. Extremely simple, Logistic Regression is often used as a baseline with respect to which more complex models are compared.

**3.3 Model evaluation**

Evaluate the model based on the following metrics:

Accuracy: Percentage of correctly classified samples.

Accuracy: The classifiers' ability to predict the correct labels for positive cases. Remember: the classifier finds all the abilities in all positive cases.

**F1 Index**: The exact and recall of the harmonic mean, reporting the model performance. Cross-validation: To prevent over-fitting, and check whether the model generalizes well we used 5-fold cross validation. We split the data into five equal parts and trained and validated the model on different subsets of the data.

**3.4 Hyperparameter tuning**

Hyperparameters of the random forest are fine-tuned using GridSearchCV. Parameters are:

**max\_depth** : maximum depth of the each tree:.

**min\_samples**\_split: Minimum number of samples needed to split an internal node. **N\_estimators**: number of trees in the forest.

After running all possible combinations, best parameters turned out to be:

Max\_deptth = 20

min\_samples\_plit = 2

n\_estimator = 300

4. Results and discussion

**4.1 Model Performance**

Use test data set to check the performance of the model. The following are the results:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| **Random Forest** | **99.31%** | **0.99** | **0.99** | **0.99** |
| **SVM** | **97.73%** | **0.98** | **0.98** | **0.98** |
| **Logistic Regression** | **96.36%** | **0.96** | **0.96** | **0.96** |

#### 

**Performance of random forest**

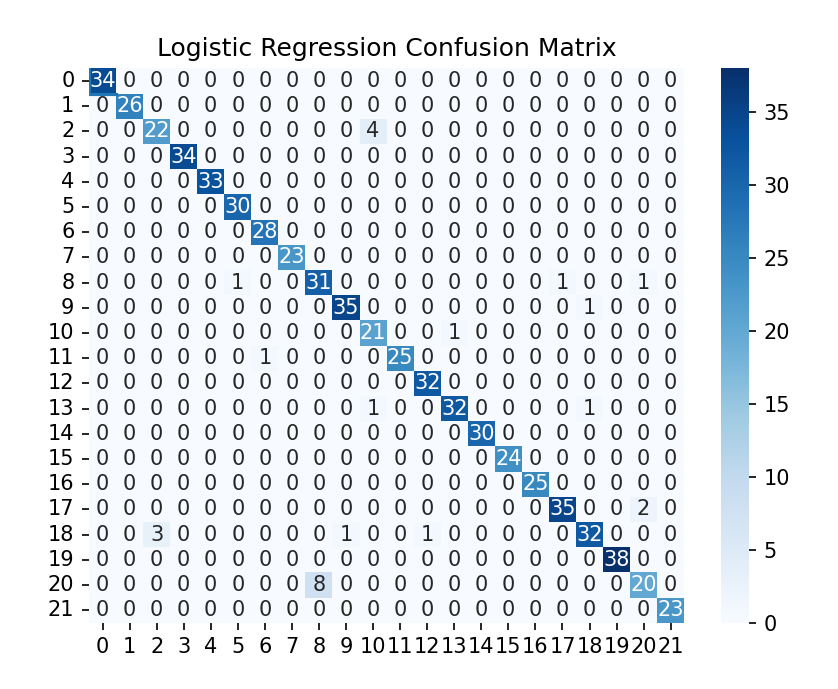
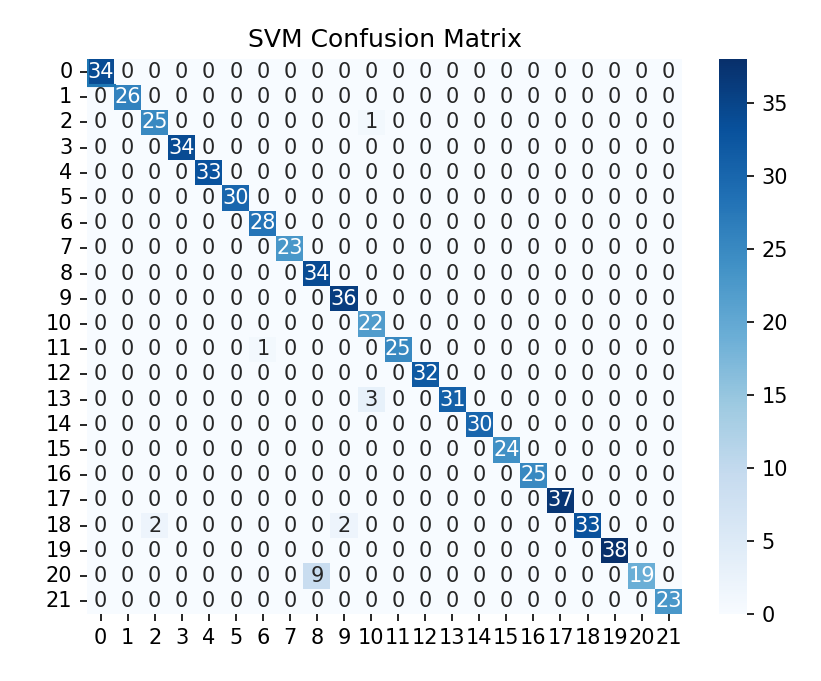
The random forest model had the maximum accuracy at 99.32%, which happens to be larger than other models; this can be contributed by its integrative nature where it aggregates predictions of many trees before arriving at a final decision. The fine-tuned model did exceptionally well on crops like apple, papaya and mango that were predicted with 100% accuracy .

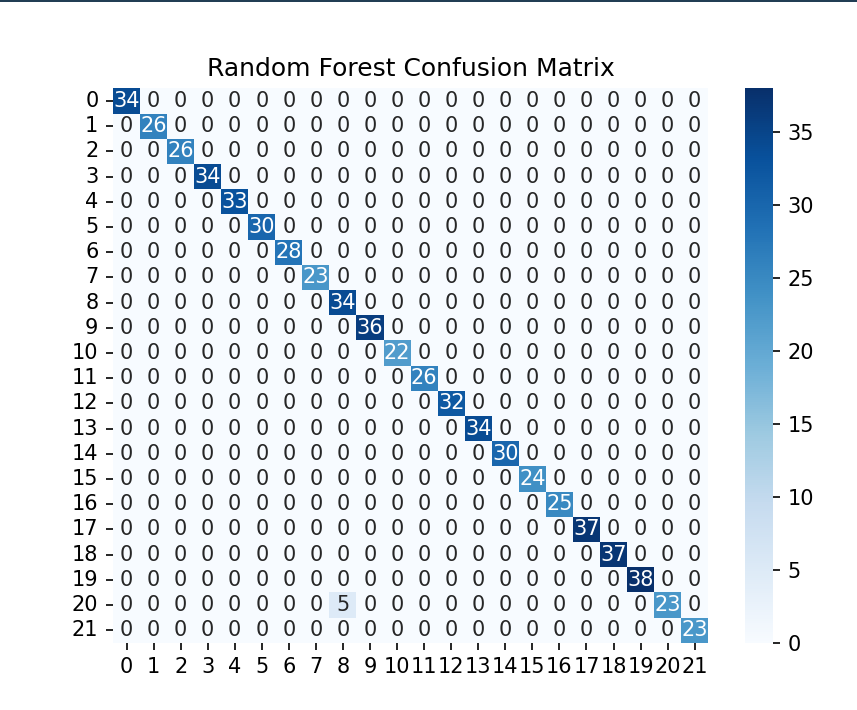
**Performance of SVM**

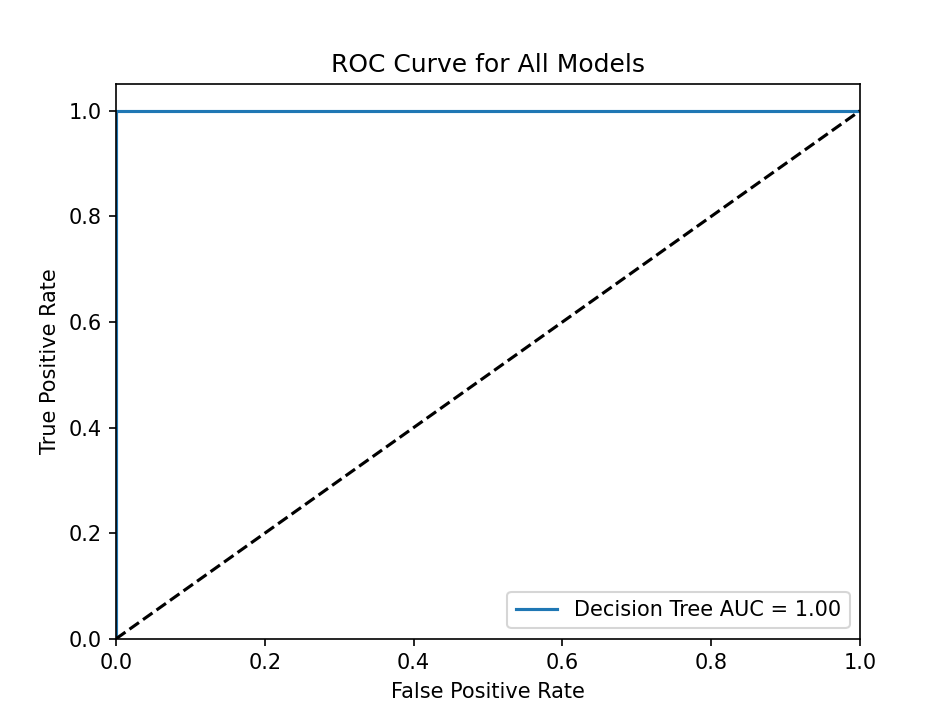
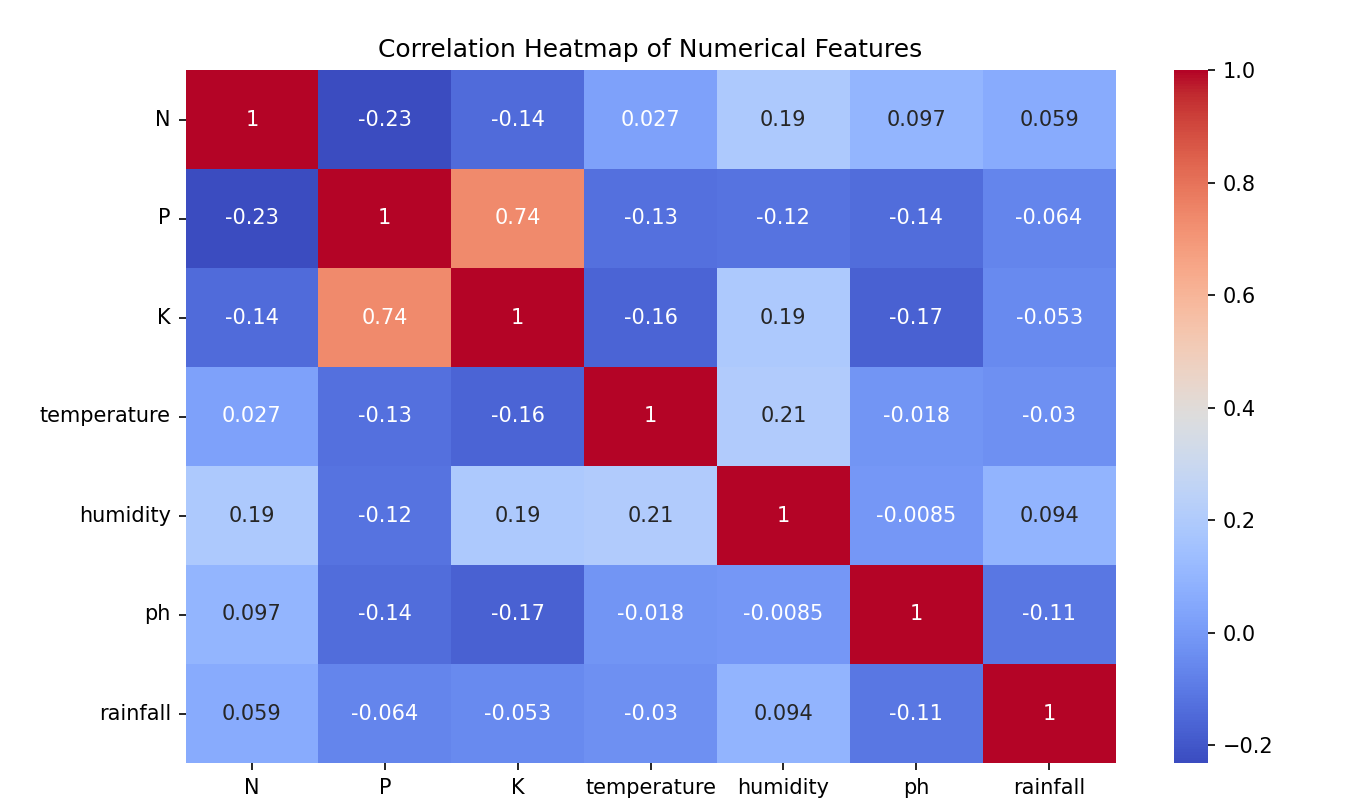
The SVM model was also tried to obtain a result of 97.73%. Although it still performs, compared to Random Forest, it was lower, and in crops such as rice, it still performs poorly, probably because different regions experience variations in temperature and humidity.

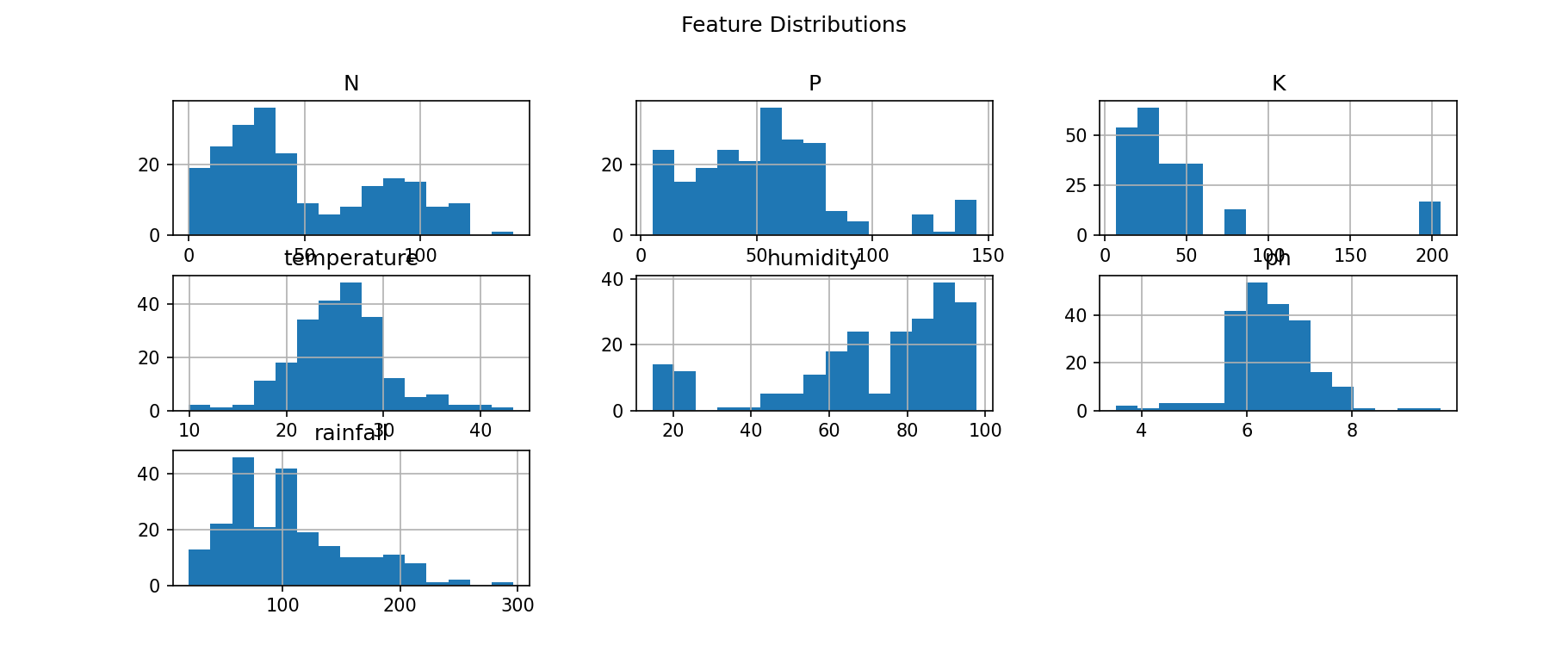
**Performance Logical Regression**

As expected, logistic regression is a simple linear classifier which achieved the lowest accuracy of 96.36 %. Although it worked satisfactorily for crops with more linear relationships between traits, like maize, it did not capture the complexity of more diverse crops like moth bean and watermelon.

****

****

****

****

**4.2. Meaning of the function**

Significance of each attribute of the Random Forest model was analyzed. The outcome showed that nitrogen (N) was the most important characteristic followed by precipitation and temperature. It shows that nutrient and environmental factors should be given significant weights on determining the most appropriate crops for a particular region. High pH importance also determines soil acidity/alkalinity as a major factor influencing crop suitability.

**5. Conclusion**

This research delves into the ability and comparison of Random Forest, SVM, and Logistic Regression in crop prediction based on soil and environmental factors. The model Random Forest, optimized for the highest accuracy, reached 99.31% and, consequently ended as the most reliable in predicting optimal crops. While slightly less accurate, it was still satisfactory compared to its adequacy, indicating that it would be quite effective. The logistic regression produced the least.

**Work for the Future:**

Further research could focus on improving yield forecasting accuracy through more integration of real-time data using IoT devices and application of deep learning algorithms involving convolutional neural networks. In combination with weather forecasting data, along with satellite images, this can enhance the sum of the model in different areas.

**6. Reference**

* **Madhuri Shripathi Rao, et al.** (2022), "Crop Prediction Using Machine Learning",Journal of Physics: Conference Series .
* **Kanaga Priya, P.**, et al. (2020), "An IoT Based Gradient Descent Approach for Precision Crop Suggestion using MLP".
* **Kalimuthu, M.**, et al. (2020), "Crop Prediction Using Machine Learning", International Conference on Smart Systems
* **Bharath S, Yeshwanth S, Yashas B L and Vidyaranya R Javalag**i 2020 Comparative Analysis of Machine Learning Algorithms in The Study of Crop and Crop yield Prediction International Journal of Engineering Research & Technology (IJERT) NCETESFT – 2020 vol 8 Issue 14.
* **Mahendra N, Vishwakarma D, Nischitha K, Ashwini and Manjuraju M. R** 2020 Crop prediction using machine learning approaches, International Journal of Engineering Research & Technology (IJERT) vol 9 Issue 8 (August 2020).
* **Gulati P and Jha S K** 2020 Efficient crop yield prediction in India using machine learning techniques International Journal of Engineering Research & Technology (IJERT) ENCADEMS – 2020 vol 8 Issue 10.
* **Gupta A, Nagda D, Nikhare P, Sandbhor A**, 2021, Smart crop prediction using IoT and machine learning International Journal of Engineering Research & Technology (IJERT) NTASU – 2020 vol 9 Issue 3.