iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73188

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Predictive Crop Selection Using Data-Driven Machine Learning Models

Patha Narasimha Rao[1], Dr. M. Dhanalakshmi[2]
[1]*Post-Graduate Student, Department of Information Technology, Data Science, Jawaharlal Nehru Technological University, Hyderabad, India*
[2]*Professor, Department of Information Technology, Jawaharlal Nehru Technological University, Hyderabad, India*

Abstract: ***With the growing need for sustainable agricultural practices, the integration of intelligent systems in farming has become increasingly essential. Crop selection, a critical decision for farmers, is often influenced by numerous environmental and soil-related factors. In this context, the present study focuses on the development of a machine learning-based crop recommendation system aimed at improving the decision-making process for crop cultivation. The system analyses various agronomic features including soil nutrients nitrogen (N), phosphorus (P), potassium (K) along with temperature, humidity, pH level, and rainfall to suggest the most appropriate crop for a given set of conditions. A dataset containing these parameters was pre-processed to remove inconsistencies and scaled using Min-Max normalization and standardization to enhance model performance. Several classification algorithms were implemented and evaluated, including Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, Gaussian Naïve Bayes, Random Forest, Gradient Boosting, AdaBoost, Bagging Classifier, and Extra Trees. These models were trained and tested using an 80-20 split of the dataset, and their performance was assessed based on accuracy metrics. Among all tested models, the Random Forest classifier emerged as the most reliable, delivering the highest prediction accuracy due to its ability to handle high-dimensional data and reduce overfitting. The system also includes a functional interface allowing users to input real-time environmental values and receive instant crop recommendations. This project demonstrates how machine learning can be effectively leveraged to support agricultural decisions, reduce crop failure risks, and enhance yield potential. By offering a data-driven approach to crop planning, the system contributes to more efficient land use, resource optimization, and long-term sustainability in agriculture.***

Keywords: Crop Recommendation, Machine Learning, Sustainable Agriculture, Soil Nutrients (NPK), Environmental Factors, Precision Farming.

I. INTRODUCTION

Agriculture has always played a fundamental role in sustaining human life and economic development, particularly in countries where a large portion of the population depends on farming for their livelihood. One of the most critical decisions in agriculture is selecting the appropriate crop for cultivation, as this choice directly influences productivity, profitability, and environmental sustainability. With the emergence of modern computing and the availability of large datasets, machine learning has become a powerful tool for transforming traditional agricultural practices. By leveraging data-driven approaches, farmers can make informed and precise decisions that are tailored to their local environmental conditions.

This project presents a machine learning-based crop recommendation system designed to assist in the scientific selection of suitable crops by analysing key agronomic factors. The system uses a diverse set of input features, including essential soil nutrients (Nitrogen, Phosphorus, and Potassium), environmental attributes (temperature, humidity, and rainfall), and soil pH level. These parameters are fed into various machine learning models that learn from historical crop production data. The dataset used for training and evaluation contains labelled records of different crops associated with specific ranges of these features. To build a robust prediction framework, multiple classification algorithms were explored and compared, including Logistic Regression, Support Vector Machines (SVM), Gaussian Naïve Bayes, Decision Trees, K-Nearest Neighbors (KNN), Random Forest, Bagging, Boosting, and others. Each model was evaluated based on its accuracy in predicting the correct crop from the test dataset. Among all the models tested, the Random Forest Classifier demonstrated the highest accuracy and reliability, making it the preferred model for final deployment.

In the proposed system, once the user inputs the required environmental and soil parameters, the model processes this information using standardized and normalized data transformations. The trained Random Forest model then predicts the most suitable crop for cultivation in those conditions. This prediction mechanism can be incorporated into a user-friendly interface for real-time use by farmers and agricultural advisors.

The implementation of such a recommendation system is a step forward in integrating artificial intelligence into agriculture. It has the potential to enhance crop yield, reduce resource wastage, and support sustainable farming by aligning crop choices with precise field conditions. Furthermore, the adaptability of the model makes it useful for diverse agricultural zones, enabling better planning and resilience in the face of climate variability.

II. LITERATURE SURVEY

A. A Review on Data Mining Techniques for Fertilizer Recommendation (2018):

Authors: Jignasha M. Jethva, Nikhil Gondaliya, Vinita Shah.

The paper explores the application of data mining techniques to enhance fertilizer recommendation strategies in agriculture. In India, farmers often apply fertilizers based on estimation, which can lead to underuse or overuse, negatively affecting crop productivity. This study reviews different data mining methodologies implemented on soil datasets to provide more precise and effective fertilizer recommendations.

B. Machine Learning: Applications in Indian Agriculture (2016):

Authors: Karandeep Kaur

This study examines the role of machine learning in addressing challenges within Indian agriculture, a sector traditionally slow to adopt technological advancements. It emphasizes how machine learning has outperformed conventional computational methods and enhanced the accuracy of AI systems, such as sensor-based tools in precision farming. The paper outlines various machine learning applications in agriculture and discusses how these technologies can help resolve the difficulties faced by Indian farmers.

C. Support Vector Machine-Based Classification Scheme of Maize Crop (2017):

Authors: Suhas S Athani, CH Tejeshwar

This paper presents a system designed to automate the identification and classification of weeds in maize crops, aiming to reduce manual labor. Using image data of maize fields, texture features are extracted and analysed through a Support Vector Machine (SVM) classifier. The model achieves an accuracy of 82% in distinguishing weeds from crops. The study also opens avenues for further research into advanced feature extraction techniques.

III.EXISTING WORK

Accurate price forecasting in agriculture often demands significant computational resources and detailed datasets, which are typically lacking in many developing nations. As a result, researchers tend to use simplified, more efficient models to meet their forecasting needs. A widely used approach in this context is time series analysis, where historical data of a variable is examined to build models that capture its patterns and trends. Significant advancements have been achieved over time in improving and optimizing time series forecasting methods. These models are advantageous because they require relatively minimal data and can deliver timely price predictions. However, to enhance the accuracy and adaptability of forecasting, there is a growing need for more robust classification systems—potentially through the use of ensemble or hybrid models that combine the strengths of multiple methods.

A. Limitations

- Many existing solutions are hardware-dependent, making them expensive to maintain and difficult for small-scale farmers to adopt.
- Although several crop recommendation tools have been developed, most still face challenges in terms of user accessibility and ease of use.
- There is a noticeable lack of innovation, with many systems repeating similar approaches without significant improvements or enhancements.

IV.PROPOSED WORK

The proposed system, titled "Crop Recommendation System Using Machine Learning Algorithms," utilizes a comprehensive dataset containing values for Nitrogen, Phosphorus, Potassium, pH, moisture, and precipitation. By incorporating a wider range of agricultural parameters compared to traditional systems, this model aims to achieve higher accuracy in crop prediction. The dataset was compiled from both direct inputs from farmers and open-source platforms such as Kaggle and GitHub, resulting in a more diverse and informative dataset.

Unlike earlier models that considered only a limited number of crop types, our system recommends a broader variety of important crops, including rice, maize, chickpea, mango, banana, mung bean, kidney bean, and apple. These crops were chosen based on their relevance across different growing seasons and agricultural zones, making the recommendations more practical and applicable. Our dataset consists of approximately 2200 data entries—significantly larger than the 1000-row datasets used in many existing systems. This expansion enhances the model's ability to learn complex patterns and generate more reliable predictions. We have implemented and tested various machine learning algorithms, such as K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Random Forest. Among these, the Random Forest classifier produced the most accurate results, especially when handling large and diverse datasets, making it the preferred algorithm for final deployment.

A. *Objectives*

1) To develop a machine learning-based system that can recommend suitable crops based on multiple environmental and soil parameters.
2) To analyze key agricultural factors such as nitrogen (N), phosphorus (P), potassium (K), pH level, temperature, humidity, and rainfall for accurate prediction.
3) To assist farmers in making data-driven decisions, moving away from traditional methods based on assumptions or peer influence.
4) To increase agricultural productivity by matching crops with the most suitable soil and climate conditions.
5) To reduce crop failure risks by using predictive models trained on real agricultural data.

V. SYSTEM ARCHITECTURE

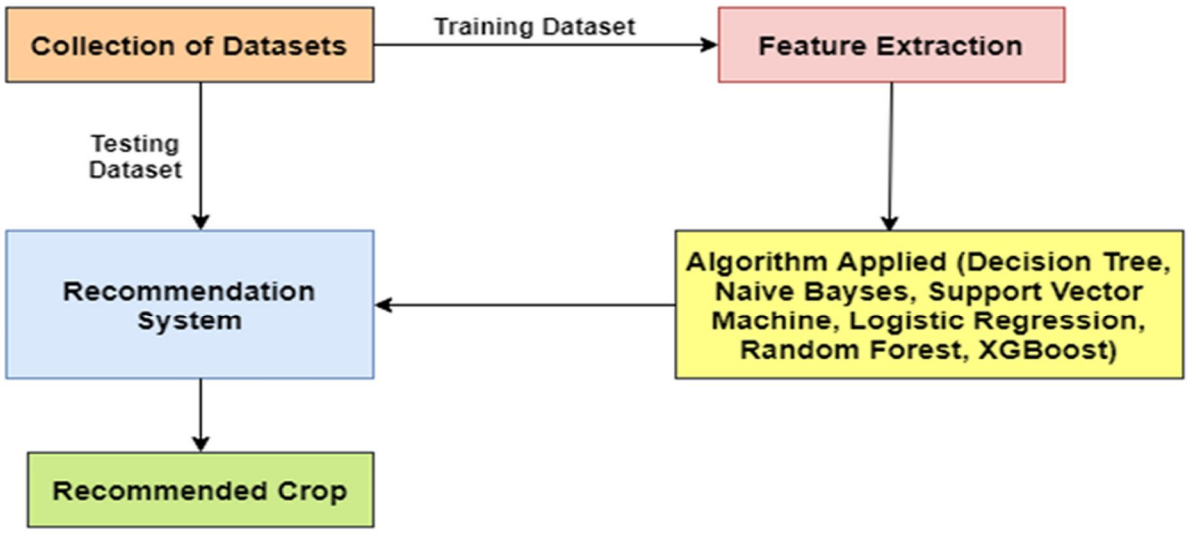


Fig 1 : Block Diagram of Overall Methodology of Proposed System

A. *Workflow*

In our framework, we have proposed a procedure that is separated into various stages as appeared in Figure 1.
The five phases are as per the following:

- Collection of Datasets
- Pre-processing (Noise Removal)
- Feature Extraction
- Applied Various Machine Learning Algorithm
- Recommendation System
- Recommended Crop

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

VI. METHODOLOGY AND IMPLEMENTATION

The implementation phase focuses on developing a machine learning model to recommend suitable crops based on soil and environmental parameters. Using a structured dataset, we preprocess the data, explore its characteristics, and apply various classification algorithms. The goal is to identify the most accurate model for predicting the best crop, ensuring efficiency and reliability in agricultural planning.

A. Data Collection

The dataset used in this study includes agricultural parameters such as Nitrogen (N), Phosphorus (P), Potassium (K), soil pH, humidity, temperature, and rainfall. It contains 2,200 historical data entries and was sourced from Kaggle. The dataset spans 22 different crops, including grains, pulses, fruits, and commercial crops such as rice, maize, chickpea, kidney beans, pigeon peas, moth beans, mung beans, black gram, lentils, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee.

B. Pre-Processing (Noise Removal):

To ensure effective implementation, data preprocessing is an essential step. Since the collected data may originate from various sources, it often comes in an unstructured or raw format, possibly containing missing values, duplicates, or inconsistencies. Therefore, this phase involves cleaning the data by eliminating redundant entries and handling incomplete information. Additionally, normalization is applied to bring all features onto a common scale for better model performance.

C. Feature Extraction

This stage emphasizes selecting the most significant features from the dataset. By doing so, any irrelevant or duplicate information is eliminated, allowing classifiers to operate more efficiently and accurately during the model training process.

D. Methodology

In the proposed system, various machine learning algorithms have been implemented, including Decision Tree, Naïve Bayes, Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and XGBoost, to evaluate and enhance crop prediction accuracy.

1) Decision Tree

The Decision Tree classifier follows a greedy approach and is a type of supervised learning algorithm. It builds a model that predicts the value or category of a target variable based on a set of decision rules derived from training data. The tree structure consists of two key components: decision nodes and leaf nodes. Decision nodes represent tests on specific features, while leaf nodes indicate the predicted outcomes. Each node evaluates an attribute, and the branches represent the possible responses to that test. This process continues recursively, constructing sub-trees for each new decision node until the final predictions are made. We have applied Decision tree approach in our model as:

(i) Importing library DecisionTreeClassifier from sklearn. tree Class
(ii) Now we create DecisionTree Classifier object
(iii) In the last we fit our data

```
# Decision Tree
from sklearn. Tree import DecisionTreeClassifier
Decision Tree=DecisionTreeClassifier (criterion="entropy", random_state=2, max_depth=5)
DecisionTree.fit (Xtrain, Ytrain)
```

2) Naïve Bayes (NB)

Naive Bayes is a classification algorithm suited for both binary and multi-class problems. It works particularly well with categorical or discrete input values due to its simplicity. The core assumption of Naive Bayes is that the presence of one feature in a class is independent of the presence of other features. This method is based on Bayes' Theorem and is especially effective when working with high-dimensional datasets. Naive Bayes is widely used in various applications such as real-time prediction, spam detection, and recommendation systems when combined with collaborative filtering. Initially, it calculates the prior probability of each class, and then the conditional probability of input features given each class to make predictions.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

We have applied Naïve Bayes (NB) approach in our model as:

(i) Importing library GaussianNB Classifier from sklearn. naive Bayes Class
(ii) Now we create Naïve Bayes Classifier object
(iii) In the last we fit our data

```
# Naive Bayses (NB)
from sklearn. naive Bayes import GaussianNB
Naive Bayes = GaussianNB ()
NaiveBayes.fit (Xtrain, Ytrain)
```

3) Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm or model which can be utilized for classification and as well as for regression challenges. However, we mainly use it in classification challenges. SVM is generally represented as training data points in space which is divided into groups by intelligible gap which is as far as possible [22]. In SVM algorithm, each data item is plotted as a point in n-dimensional space with each feature value being the value of a specific coordinate. Then the classification is performed by finding the hyper-plane differentiating the two classes very well.

We have applied Support Vector Machine (SVM) approach in our model as:

(i) Importing library SVC from sklearn.svm Class
(ii) Now we create SVM classification object
(iii) At last, we fit our data

```
# Support Vector Machine (SVM)
from sklearn.svm import SVC
SVM = SVC (gamma='auto')
SVM.fit (Xtrain, Ytrain)
```

4) Logistic Regression (LR)

Logistic Regression is a widely used statistical technique designed to model a binary outcome using a logistic (sigmoid) function. In its basic form, it estimates the relationship between input features and a binary dependent variable. Although it is commonly applied to binary classification tasks, more advanced variants extend its capability to multi-class problems. Unlike linear regression, logistic regression predicts the probability of class membership, making it suitable for classification rather than continuous value prediction.

We have applied Logistic Regression (LR) in our model as:

(i) Importing library LogisticRegression from sklearn. Linear Class
(ii) Now we create LogReg classifier object
(ii) In the last we fit our data

```
# Logistic Regression
from sklearn. linear model import LogisticRegression
LogReg = LogisticRegression(random_state=2)
LogReg.fit (Xtrain, Ytrain)
```

5) Random Forest (RF)

Random Forest is a machine learning algorithm that builds multiple decision trees during training. For classification tasks, it outputs the class with the majority vote, and for regression, it predicts the average of outputs. The model's accuracy generally improves as the number of trees increases. In this system, training is performed using features such as rainfall, precipitation, temperature, and crop yield. About two-thirds of the dataset is used for training, while the remaining portion is reserved for testing and evaluation. The Random Forest algorithm relies on three key parameters:

- **n_estimators (n_tree):** The number of trees to be grown in the forest.
- **max_features (m_try):** The number of features considered when splitting a node.
- **min_samples_leaf (node size):** The minimum number of samples required at a leaf node.

We have applied Random Forest (RF) in our model as:

(i) Importing library RandomForestClassifier from sklearn. ensemble Class
(ii) Now we create RF classifier object
(iii) In the last we fit our data

```
# Random Forest
from sklearn. ensemble import RandomForestCl assifier
RF = RandomForestClassifier (n_estimators=20, random_state=0)
RF.fit (Xtrain, Ytrain)
```

6) XGBoost

XGBoost, short for eXtreme Gradient Boosting, is an enhanced and efficient version of the traditional gradient boosting algorithm. It is designed to deliver high performance, speed, and accuracy in model training. Recognized for outperforming many other machine learning models, XGBoost is widely used in data science competitions and practical applications. It is an open-source tool and a key component within the Distributed Machine Learning Community. By implementing parallel tree boosting, also known as Gradient Boosted Decision Trees (GBDT), XGBoost efficiently handles a variety of predictive modeling tasks.

We have applied XGBoost in our model as:

(i) Importing library xgboost
(ii) Now we create XB classifier object
(iii) In the last we fit our data

```
# XGBoost
import xgboost as xgb
XB = xgb. XGBClassifier ()
XB.fit (Xtrain, Ytrain)
```

VII. PERFORMANCE EVALUATION

The crop recommendation system was tested using ten different machine learning models. These included Logistic Regression, Naive Bayes, SVM, KNN, Decision Tree, Random Forest, and others. The dataset was split 80/20 for training and testing, with normalization and standardization applied. Among all models, Random Forest achieved the highest accuracy and was selected for the final prediction model due to its strong and reliable performance.

A. Model Performance Evaluation

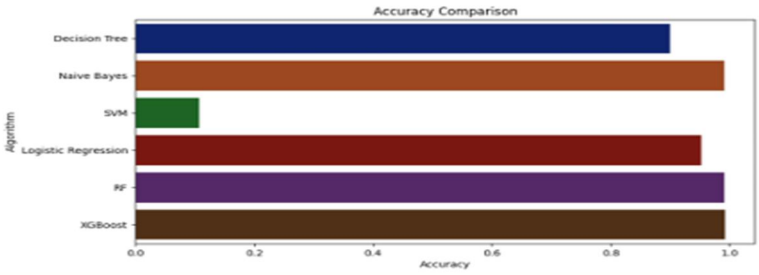


Fig 2 : Accuracy Comparison

```
Decision Tree --> 0.9
Naive Bayes --> 0.990909090909091
SVM --> 0.10681818181818181
Logistic Regression --> 0.9522727272727273
RF --> 0.990909090909091
XGBoost --> 0.9931818181818182
```

Fig 3: Machine Learning vice Accuracy Result

VIII. RESULTS

It highlights the accuracy of each model, compares their effectiveness, and identifies the most reliable approach based on predictive performance. These results help in determining the best-suited model for real-world crop recommendation tasks. Below shows the example results. The user provides the input data like Temperature, Humidity, PH, Rainfall, Potassium (K) in soil, Nitrogen (N) in soil, Phosphorous (P) in soil.

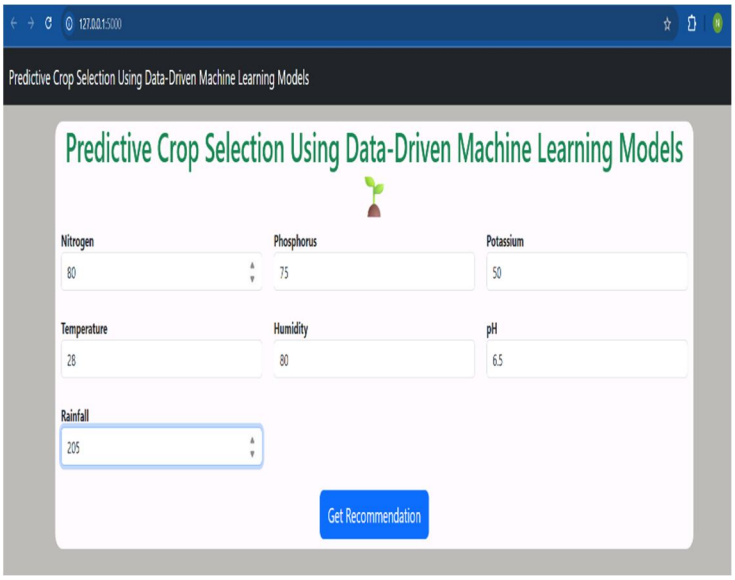


Fig 4: Giving inputs in the webpage

After providing the input click “Get Recommendation” and then user will get the output as follows.

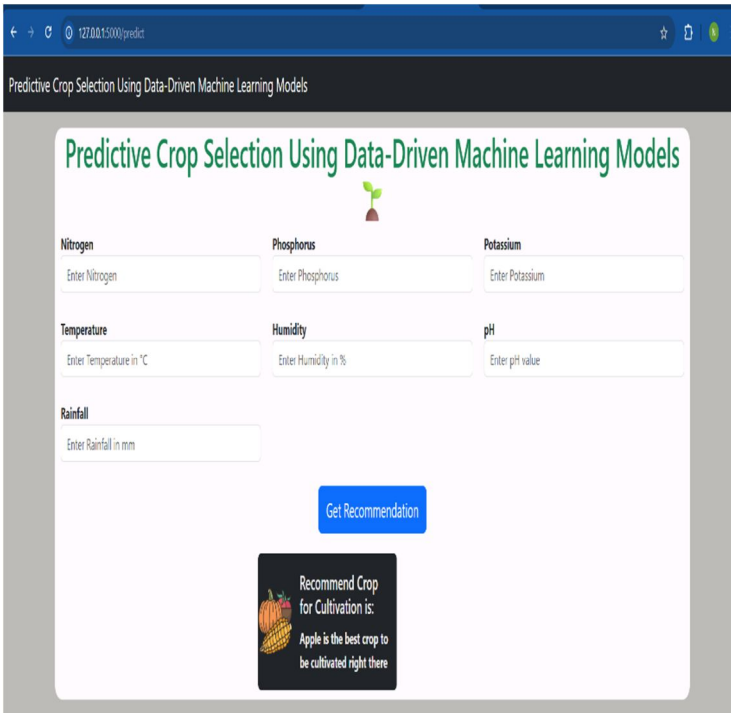


Fig 5: Output Display

Based on the given input parameters output will be displayed.

IX. CONCLUSION

This paper introduces an advanced crop recommendation system designed to assist farmers throughout India in making well-informed decisions on which crops to grow. The system evaluates several environmental and soil factors such as Nitrogen, Phosphorus, Potassium, pH level, temperature, humidity, and rainfall. By utilizing this approach, farmers can select the most suitable crop, potentially increasing their yields and contributing to the country's agricultural productivity and economic growth. The study evaluates the performance of six different machine learning algorithms—Decision Tree, Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest, and XGBoost—in providing crop recommendations. Among these models, XGBoost emerged as the most effective, delivering the highest accuracy in predictions.

A. Future Scope

The system can be enhanced further to add following functionality:

1) The main future work's aim is to improved dataset with larger number of attributes.
2) We need to build a model, which can classify between healthy and diseased crop leaves and also if the crop has any disease, predict which disease is it.
3) To build website and mobile app for easy to use.
4) Integrating deep learning to allow users to upload images of soil or plants for diagnosing diseases or nutrient deficiencies.
5) To serve a wider audience, particularly in India, the application can be localized into regional languages such as Hindi, Telugu, Tamil, Kannada, etc.
6) Provide offline support using cached data for areas with limited internet connectivity, enabling the tool to be useful even without real-time access.

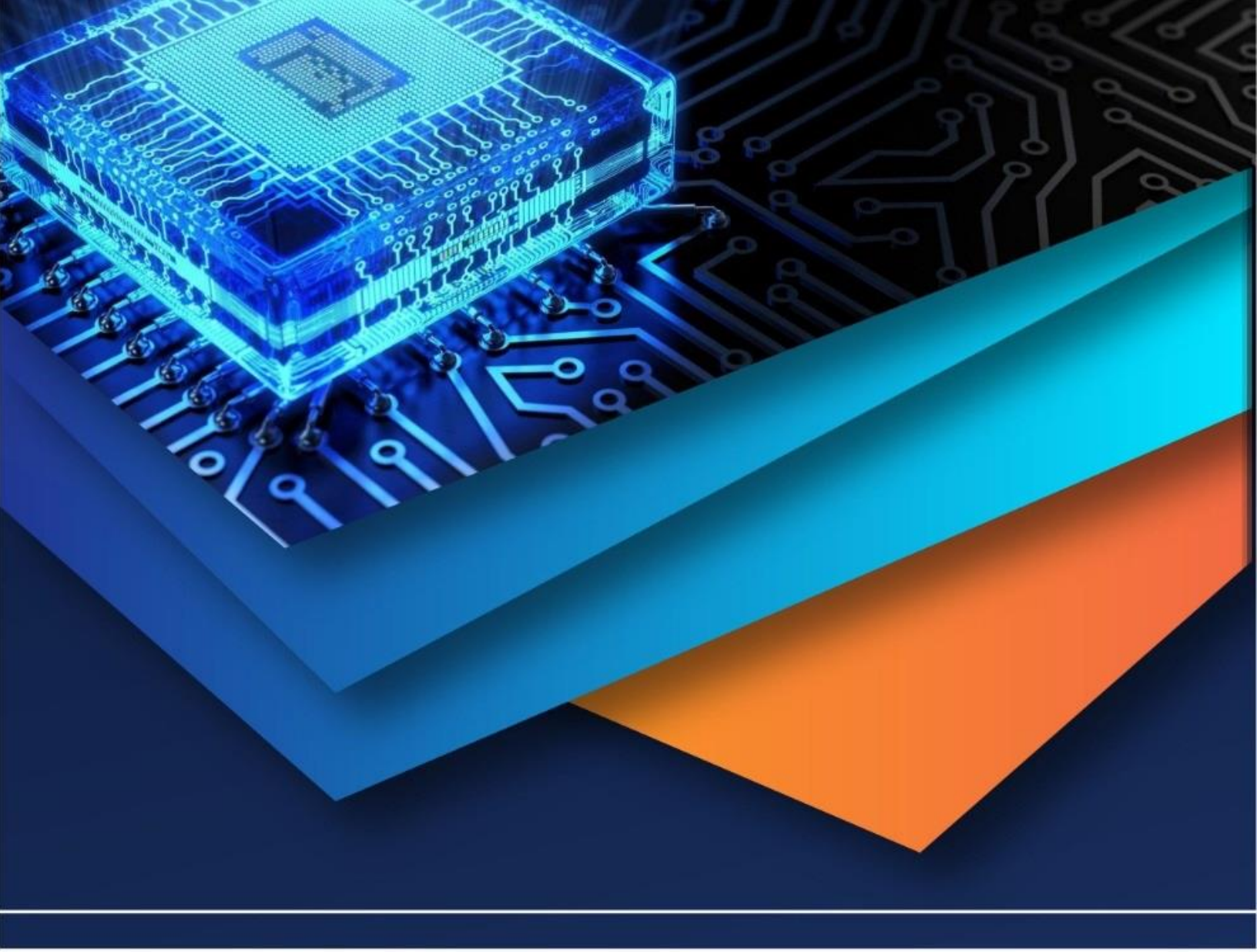
X. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who supported me throughout the completion of this project, "Predictive Crop Selection Using Data-Driven Machine Learning Models."

First and foremost, I extend my heartfelt thanks to my project guide, Dr. M. Dhanalakshmi, for their valuable guidance, encouragement, and constructive suggestions throughout the course of this project. Their continuous support and insightful feedback greatly enhanced the quality of this work. I am also thankful to the faculty members of the Department of Information Technology, JNTUH, for their academic support and encouragement. Finally, I am deeply grateful to my family for their constant support, motivation, and encouragement, which enabled me to complete this project successfully.

REFERENCES

[1] Kumar, Y. Jeevan Nagendra, V. Spandana, V. S. Vaishnavi, K. Neha, and V. G. R. R. Devi. "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector." In 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 736-741. IEEE, 2020.
[2] Nigam, Aruvansh, Saksham Garg, Archit Agrawal, and Parul Agrawal. "Crop yield prediction using machine learning algorithms." In 2019 Fifth International Conference on Image Information Processing (ICIIP), pp. 125-130. IEEE, 2019.
[3] Kulkarni, Nidhi, H, G.N. Srinivasan, B.M. Sagar, and N.K. Cauvery. "Improving Crop Productivity through a Crop Recommendation System using Ensemble technique" in 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), pp. 114-119. IEEE, 2018.
[4] Rajak, Rohit Kumar, Ankit Pawar, Mitlee Pendke, Pooja Shinde, Suresh Rathod and Avinash Devare, "Crop recommendation system to maximise crop yield using machine learning technique". International Research Journal of Engineering and Technology 4, no. 12 (2017): 950-953.
[5] Medar, Ramesh, Vijay S. Rajpurohit, and Shweta. "Crop yield prediction using machine learning techniques." In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1-5. IEEE, 2019.
[6] Kamatchi, S. Bangaru, and R. Parvathi. "Improvement of Crop Production Using Recommender System by Weather Forecasts." Procedia Computer Science 165 (2019): 724-732.
[7] Kulkarni, Nidhi H., G. N. Srinivasan, B. M. Sagar, and N. K. Cauvery. "Improving Crop Productivity Through Recommendation System A Crop Using Ensembling Technique." In 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), pp. 114-119. IEEE, 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY