JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Crop Prediction Based on Soil pH and Weather Conditions

Srinidhi V, Uppaluru SVN Santhosha Roopa, Yashitha U Raithatha, Anushka Thakur, Samudrala SowmyaSree

Bhavan's Vivekananda College of Science, Humanities and Commerce

Research Guided by Ms. B Divya Rekha (Asst. Professor)

Email: [Divya Rekha, B.(2025),personal contact] b.d.rekha0310@gmail.com

Abstract

In India, farming is crucial for both making a living and ensuring there is enough food. However, farmers are facing more problems due to unpredictable weather and poor soil health. It's important to choose the right crops to plant, but traditional methods based on past experience often aren't enough. This research looks at how using data and a method called logistic regression can help. Logistic regression is a simple but powerful tool that predicts the best crops to plant by looking at factors like weather and soil quality. We use data such as temperature, rainfall, soil pH, and nutrient levels to make predictions. The results are encouraging. Our model not only finds the key factors affecting crop growth but also makes better predictions than older methods. This means farmers and planners can make smarter decisions, even when the environment is uncertain. In the end, this study shows how combining modern technology with farming can tackle major global issues. By using these predictive tools, we can move towards sustainable farming, better productivity, and improved food security.

Keywords: Logistic regression, Predictive analytics, Crop selection, Soil health, Climate impact, Sustainable agriculture, Food security.

Introduction

Background

Food growth is very important and helps people live and advance, especially in India. India's agricultural story is long and amazing, which makes it one of the best places to grow crops and produce food. More than 31% of people work in agriculture there, and it brings 15.4% to the pile of country money. But there are problems with which it's hard to deal with climate change, strange patterns of rain, and to make enough food for everyone without much.

Statement

Farmers often struggle with choosing the right crops, optimizing yields, and managing resources like soil and fertilizers. Traditional farming practices can fall short in addressing these challenges. This is where technology, especially machine learning, can help. By using data-driven insights, farmers can make better decisions about crop selection, resource management, and timing, improving productivity and profitability.

Objectives

- To improve crop yield predictions using machine learning techniques such as Logistic Regression and Linear Regression.
- To compare these models using metrics like mean absolute error to determine the best approach for aiding farmers in decision-making.
- To provide accurate predictions that help farmers adapt to changing conditions, optimize resources, and increase yields sustainably.

Scope of the Project

This study aims to connect traditional agriculture with modern technology, helping farmers handle uncertainties and succeed. By integrating machine learning into farming, the project seeks to enhance productivity and contribute to a sustainable and resilient agricultural future for India and beyond. The project

will focus on analyzing data on soil type, temperature, rainfall, and fertilizer needs to provide data-driven insights for better farming practices.

Design Architecture

Logistic Regression

Logistic regression be a machine learning that helps computers make decisions in two-choice situations. What it does is predict if something belongs to one group or the other. It predicts the likelihood that a court belongs to a specific class using the sigmoid function. The answer should always be between 0 and 1. so, if the result comes out more than 0.5, it's categorized as Class 1. If it not, it gets placed in Class 0.

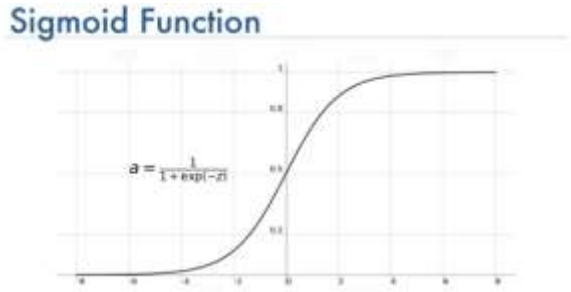
***Key Points**:*

- Predicts categorical outcomes (e.g., Yes/No, 0/1).
- Produces values which are random between 0 and 1.
- Fits an "S" shaped logistic function instead of a regression line.

Logistic Function – Sigmoid Function

The sigmoid function maps predicted values to probabilities, producing an "S" shaped curve that ranges between 0 and 1.

Fig:1.0



OneStep-elecTRON. (n.d.). *Sigmoid Function* [Image]. GitHub.

How does Logistic Regression Work?

1. Data: Collect data with features (X) and outcomes (Y) where Y is binary.

2. Model: Take your inputs and apply a multi-linear function to it using the sigmoid function to get the plausible results.

3. Sigmoid Function: Transforms continuous values into probabilities between 0 and 1.

Decision tree classification

Understanding Decision Tree

In order to make decisions easily we can opt out for decision tree which is a simple diagram that shows varied choices and their feasible results. it's a visual representation of choices for solving a problem which shows how varied factors are related.

Decision Tree Structure

- Root Node: It is the origin that stands for the entire set of information.
- Branches: Lines connecting nodes, showing the flow from one decision to another.
- Internal Nodes: These are the areas where decisions are based on input attributes.
- Leaf Nodes: they are very important terminal nodes which exist at the end of the branches and which also represents concluding outcomes or predictions.

They help with decision-making by visualizing outcomes, making it easier to compare and evaluate different options. Example: Deciding whether to drink coffee based on the time of day and tiredness.

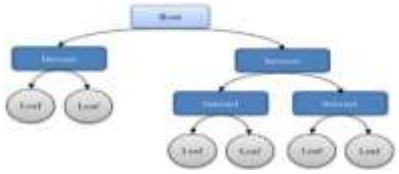


Fig:1.1

PeerJ Computer Science. (n.d.). *Sample structure of a decision tree* [Figure]. ResearchGate.

How Decision Trees Work

- Root Node: Starts with a main question based on dataset features.
- Branches: Ask a series of yes/no questions to split data into subsets.
- Final Outcome: Follow branches to reach the end, where the final decision or prediction is made.

Advantages of Decision Trees

- Simplicity and Interpretability: Easy to understand and visualize like a flowchart.
- Versatility: Can be used for classification and regression tasks.
- No Need for Feature Scaling: No need to normalize or scale data.
- Operates erratic affiliations: expresses nonlinear affiliations between attributes and final goal variables.

Disadvantages of Decision Trees

- Overfitting: Can spot out commotion and particulars in learning data, executing on new information.
- Instability: Small input variations can lead to significant prediction differences.
- Bias towards Features with More Levels: Can focus too much on features with many categories, missing other important features.

Random Forest Model

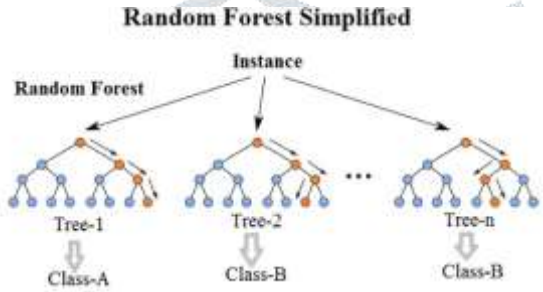
Understanding Intuition for Random Forest Algorithm

In order to make predictions, there is a powerful machine learning technique called Random Forest which uses multiple decision trees. Each tree is trained on a different random part of the dataset. The final prediction is made by combining the results from all trees, which improves accuracy. **Example**: Imagine asking a group of friends for vacation advice. Each friend gives their recommendation, and you make the final decision based on the majority opinion or averaging their suggestions.

Key Features of Random Forest

- Handles Missing Data: Automatically manages missing values during training.
- Feature Importance: Ranks features based on their importance in making predictions.
- Scales Well: Performs well with large and complex data.
- Versatility: Can be used for both classification (predicting categories) and regression (predicting continuous values).

Fig:1.2



Linedata. (n.d.). Random Forest Simplified [Image]

How Random Forest Algorithm Works

1. Build Multiple Trees: In order to do this, we use Random Forest, which helps create multiple decision trees using sample of data which are random.

2. Random Feature Selection: A arbitrary subset of features is used to break the data, adding variety for every single tree.

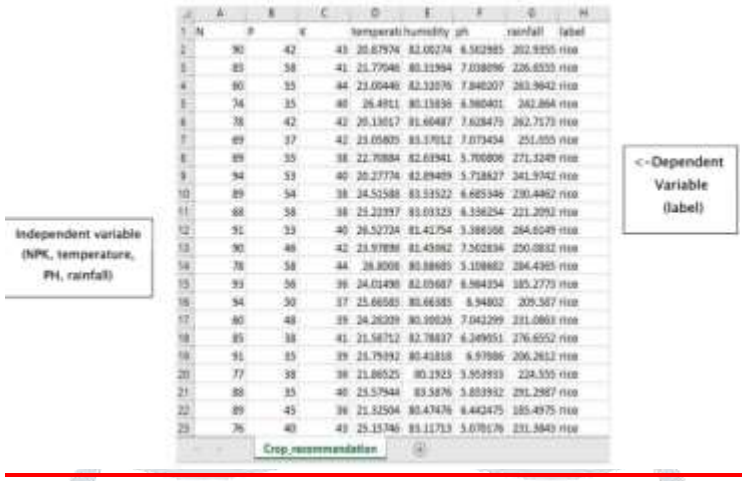
3. Independent Predictions: Based on its learning data prediction is made for each tree.

Combine Predictions:

- For classification, the final prediction is the majority vote from all trees.
- For retrogression, the concluding prediction is the standard of all tree's prognosis.

This randomness in data samples and feature selection helps prevent overfitting, making the model more accurate and reliable.

Dataset evaluation (Table:1.0)



Logistic regression (methodology)

This code trains a Logistic Regression model to predict crop types based on soil and climate features. It first encodes crop labels into numerical values using Label Encoder, then selects features like ph, N, P, K, temperature, and rainfall as inputs (X) and the encoded crop labels as targets (y). The dataset is split into training (80%) and testing (20%) using train_test_split. The model is trained with a max of 200 iterations and then used to predict crop labels for the test set. The Evaluation of the performance is done using accuracy score, a report is categorized (showing precision, recall, and F1-score for each crop), and a confusion matrix to analyses misclassifications. This approach helps in identifying the best crop for given soil and weather conditions.

Decision tree (methodology)

This code trains a Decision Tree Classifier to predict crop types while handling class imbalance using SMOTE (Synthetic Minority Over-Sampling Technique). It starts by encoding crop labels into numerical values with Label Encoder, then separates features (X) from the target labels (y). Since some crops might be underrepresented, SMOTE balances the dataset by generating synthetic samples. The balanced dataset is split into training (80%) and testing (20%) using train_test_split. A Decision Tree Classifier is trained with constraints like max_depth=10, min_samples_split=5, and min_samples_leaf=3 to prevent overfitting. The model's performance is evaluated using 5-fold cross-validation and then tested on unseen data. It calculates a precision and a categorized report, detailing accuracy, recall, and F1-score for every single crop type. This process ensures the model learns well despite class imbalances and generalizes better to new data.

Random forest (methodology)

This code trains a Random Forest Classifier to predict crop types based on given features. First, it separates the features (X) and target labels (y) from the dataset. using train_test_split method, the information is the broken into 80% training and 20% testing. To ensure all features are on the same scale, StandardScaler is applied, transforming both the training and test data. A Random Forest model with 200 decision trees (n_estimators=200) is trained on the scaled training data. The trained model then predicts crop types for the test set, and its performance is evaluated using 38 accuracy score and a classification report. Finally, the accuracy percentage is printed, showing how well the model classifies crops.

Software Used

- Python: Main programming language
- Libraries: Pandas, NumPy, Scikit-learn, imblearn (SMOTE), joblib, Matplotlib, seaborn
- Jupyter Notebook: For model development and visualization

Comparative analysis (Tables:1.1,1.2,1.3)

Logistic Regression: 0.87% **Decision tree classifier: 0.88%** **Random Forest: 0.99%**

Accuracy: 0.8681818181818182

Classification Report:

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	23
banana	1.00	1.00	1.00	21
blackgram	0.65	0.85	0.74	20
chickpea	1.00	1.00	1.00	26
coconut	1.00	0.93	0.96	27
coffee	0.89	1.00	0.94	17
cotton	0.79	0.88	0.83	17
grapes	1.00	1.00	1.00	14
jute	0.87	0.87	0.87	23
kidneybeans	0.75	0.75	0.75	20
lentil	0.62	0.91	0.74	11
maize	0.89	0.76	0.82	21
mango	0.95	1.00	0.97	19
mothbeans	0.64	0.29	0.40	24
mungbean	0.50	0.68	0.58	19
muskmelon	1.00	1.00	1.00	17
orange	0.93	1.00	0.97	14
papaya	0.88	0.91	0.89	23
pigeonpeas	0.83	0.65	0.73	23
pomegranate	1.00	0.96	0.98	23
rice	0.89	0.84	0.86	19
watermelon	1.00	1.00	1.00	19
accuracy			0.87	440
macro avg	0.87	0.88	0.87	440
weighted avg	0.87	0.87	0.86	440

Accuracy: 0.8795454545454545

Classification Report:

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	23
banana	1.00	1.00	1.00	21
blackgram	0.27	1.00	0.43	20
chickpea	1.00	0.00	0.00	26
coconut	1.00	0.00	0.00	27
coffee	1.00	1.00	1.00	17
cotton	1.00	1.00	1.00	17
grapes	1.00	1.00	1.00	14
jute	1.00	1.00	1.00	23
kidneybeans	1.00	1.00	1.00	20
lentil	1.00	1.00	1.00	11
maize	1.00	1.00	1.00	21
mango	1.00	1.00	1.00	19
mothbeans	1.00	1.00	1.00	24
mungbean	1.00	1.00	1.00	19
muskmelon	1.00	1.00	1.00	17
orange	1.00	1.00	1.00	14
papaya	1.00	1.00	1.00	23
pigeonpeas	1.00	1.00	1.00	23
pomegranate	1.00	1.00	1.00	23
rice	1.00	1.00	1.00	19
watermelon	1.00	1.00	1.00	19
accuracy			0.88	440
macro avg	0.97	0.91	0.88	440
weighted avg	0.97	0.88	0.85	440

Model Accuracy: 99.32%

Classification Report:

	precision	recall	f1-score	support
apple	1.00	1.00	1.00	23
banana	1.00	1.00	1.00	21
blackgram	1.00	1.00	1.00	20
chickpea	1.00	1.00	1.00	26
coconut	1.00	1.00	1.00	27
coffee	1.00	1.00	1.00	17
cotton	1.00	1.00	1.00	17
grapes	1.00	1.00	1.00	14
jute	0.92	1.00	0.96	23
kidneybeans	1.00	1.00	1.00	20
lentil	0.92	1.00	0.96	11
maize	1.00	1.00	1.00	21
mango	1.00	1.00	1.00	19
mothbeans	1.00	0.96	0.98	24
mungbean	1.00	1.00	1.00	19
muskmelon	1.00	1.00	1.00	17
orange	1.00	1.00	1.00	14
papaya	1.00	1.00	1.00	23
pigeonpeas	1.00	1.00	1.00	23
pomegranate	1.00	1.00	1.00	23
rice	1.00	0.89	0.94	19
watermelon	1.00	1.00	1.00	19
accuracy			0.99	440
macro avg	0.99	0.99	0.99	440
weighted avg	0.99	0.99	0.99	440

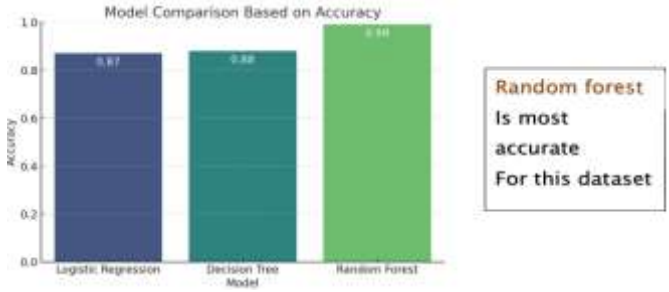
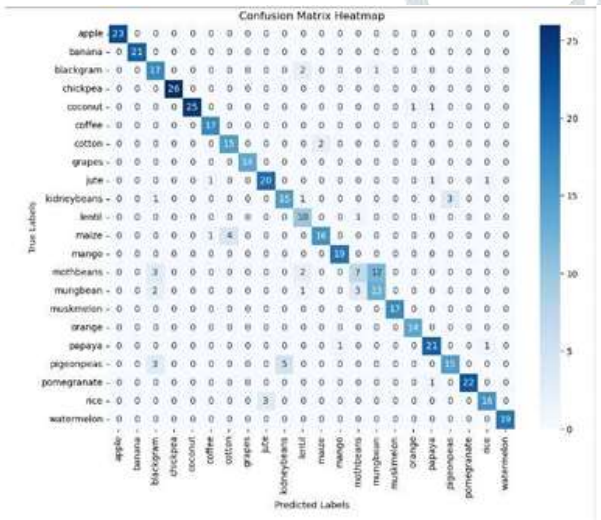


Fig:1.3

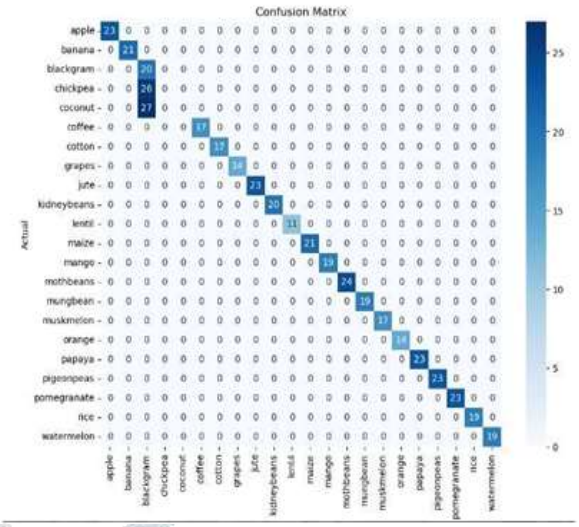
Result (Visualization)

Confusion Matrix (Logistic and Decision tree)

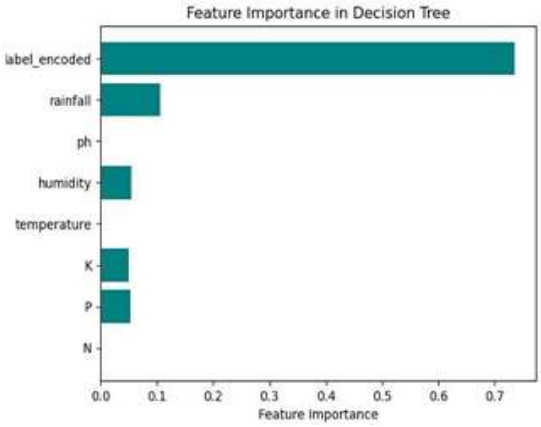
Logistic Regression (Fig:1.4)



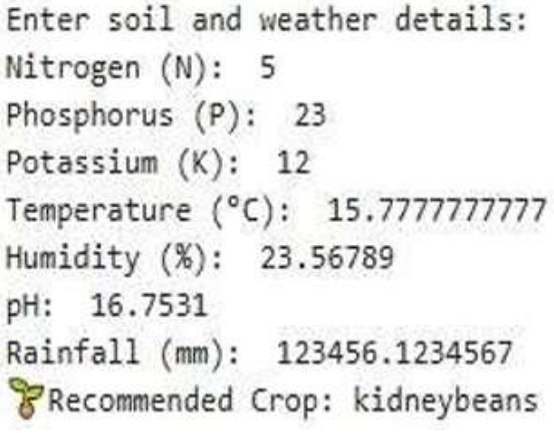
Decision Tree Classifier (Fig:1.5)



Bias Variables in Decision Tree (Fig:1.6)



Prediction Result of Random Forest (Fig:1.7)



Conclusion

This crop recommendation system helps farmers make data driven decisions by analyzing soil and weather conditions to suggest the best crops. Using Random Forest Classifier, we achieved high accuracy, ensuring reliable and practical recommendations. This system can enhance agricultural productivity, reduce trial-and-error farming, and promote sustainable farming practices.

Future Scope

1. Integration with IoT & Sensors → Real-time data collection for more accurate predictions.
2. Mobile & Web Application → Making the system accessible to farmers on their devices.
3. AI-Powered Optimization → Using Deep Learning for even better accuracy.
4. Regional Adaptations → Expanding the model for different climatic zones and soil types.
5. Water & Fertilizer Advice → Adding irrigation and fertilizer recommendations for better yield.

References

Addu, S., Sheelam, S., Mekala, S., Sulthana, N., Mekala, L., & Alsalami, Z. (2024). Assessing Environmental Impact: Machine Learning for Crop Yield Prediction. *E3S Web of Conferences*, *529*, 03008. https://doi.org/10.1051/e3sconf/202452903008

Ahmed, F. U., Das, A., & Zubair, M. (2024). A Machine Learning Approach for Crop Yield and Disease Prediction Integrating Soil Nutrition and Weather Factors. *ArXiv (Cornell University)*, 1–6. https://doi.org/10.1109/icaccess61735.2024.10499459

Anne Marie Chana, Bernabé Batchakui, & Boris Bam Nges. (2023). Real-Time Crop Prediction Based on Soil Fertility and Weather Forecast Using IoT and a Machine Learning Algorithm. *Agricultural Sciences*, *14*(05), 645–664. https://doi.org/10.4236/as.2023.145044

Dey, B., Ferdous, J., & Ahmed, R. (2024). Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. *Heliyon*, e25112–e25112. https://doi.org/10.1016/j.heliyon.2024.e25112

Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., Shdefat, A., & Saker, L. (2023). Crop Prediction Model Using Machine Learning Algorithms. *Applied Sciences*, *13*(16), 9288. https://doi.org/10.3390/app13169288

GUPTA, R., SHARMA, A., GARG, O., MODI, K., & KASIM, S. (2021, October 4). *WB-CPI: Weather Based Crop Prediction in India Using Big Data Analytics | IEEE Journals & Magazine | IEEE Xplore*. Ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/9557312

Lidbe, P., Gajbhiye, M., Meshram, S., Deogade, M., Belorkar, P., & Bhagyashree Dharaskar, P. (2024). GEOLOCATION BASED CROP PREDICTION SYSTEM. *INTERNATIONAL JOURNAL of PROGRESSIVE RESEARCH in ENGINEERING MANAGEMENT and SCIENCE (IJPREMS)*, *04*, 1404–1411. https://www.ijprems.com/uploadedfiles/paper/issue_5_may_2024/34373/final/fin_ijprems1716621206.pdf

Mahendra , N., Dhanush, V., Nischitha, K., Ashwini, M., & Manjuraju, M. R. (2020). Crop Prediction using Machine Learning Approaches. International Journal of Engineering Research And, V9(08). https://doi.org/10.17577/ijertv9is080029

Mirpulatov, I., Gasanov, M., & Matveev, S. (2023). Soil Dynamics and Crop Yield Modeling Using the MONICA Crop Simulation Model and Time Series Forecasting Methods. *Agronomy*, *13*(8), 2185. https://doi.org/10.3390/agronomy13082185

Pierre, N., Ishimwe Viviane Ishimwe Viviane, Lambert, U., Ishimwe Viviane, Irakora Shadrack, Bakunzi Erneste, Nshimyumuremyi Schadrack, Alexis, N., Francois, K., & Habiyaremye Theogene. (2023). AI Based Real-Time Weather Condition Prediction with Optimized Agricultural Resources. *AI Based Real-Time Weather Condition Prediction with Optimized Agricultural Resources*, *7*(2), 36–49. https://doi.org/10.47672/ejt.1496

Purohit, K., Kumar Singh, A., & Chatterjee, S. (2024). Enhancing agriculture production through smart assessment of soil nutrients. *Journal of Crop Improvement*, *38*(4), 392–409. https://doi.org/10.1080/15427528.2024.2355249

Rani, S., Mishra, A. K., Kataria, A., Mallik, S., & Qin, H. (2023). Machine learning-based optimal crop selection system in smart agriculture. *Scientific Reports*, *13*(1), 15997. https://doi.org/10.1038/s41598-023-42356-y

Rizvi, C. M., Singh, Er. S., & Kumar, A. (2024). Predictive Analytics for Better Crop Management and Production using Machine Learning. *Emerging Trends in IoT and Computing Technologies*, 41–46. https://doi.org/10.1201/9781003535423-8

Sharma, P., Pankaj Dadheech, & A.V. Senthil Kumar. (2023). AI-Enabled Crop Recommendation System Based on Soil and Weather Patterns. *Practice, Progress, and Proficiency in Sustainability*, 184–199. https://doi.org/10.4018/978-1-6684-8516-3.ch010

Singh, D., Sobit, R., & Kumar Malik, P. (2023). Retracted: IoT-Driven Model for Weather and Soil Conditions Based on Precision Irrigation Using Machine Learning. *Security and Communication Networks*, *2023*, 1–1. https://doi.org/10.1155/2023/9849621