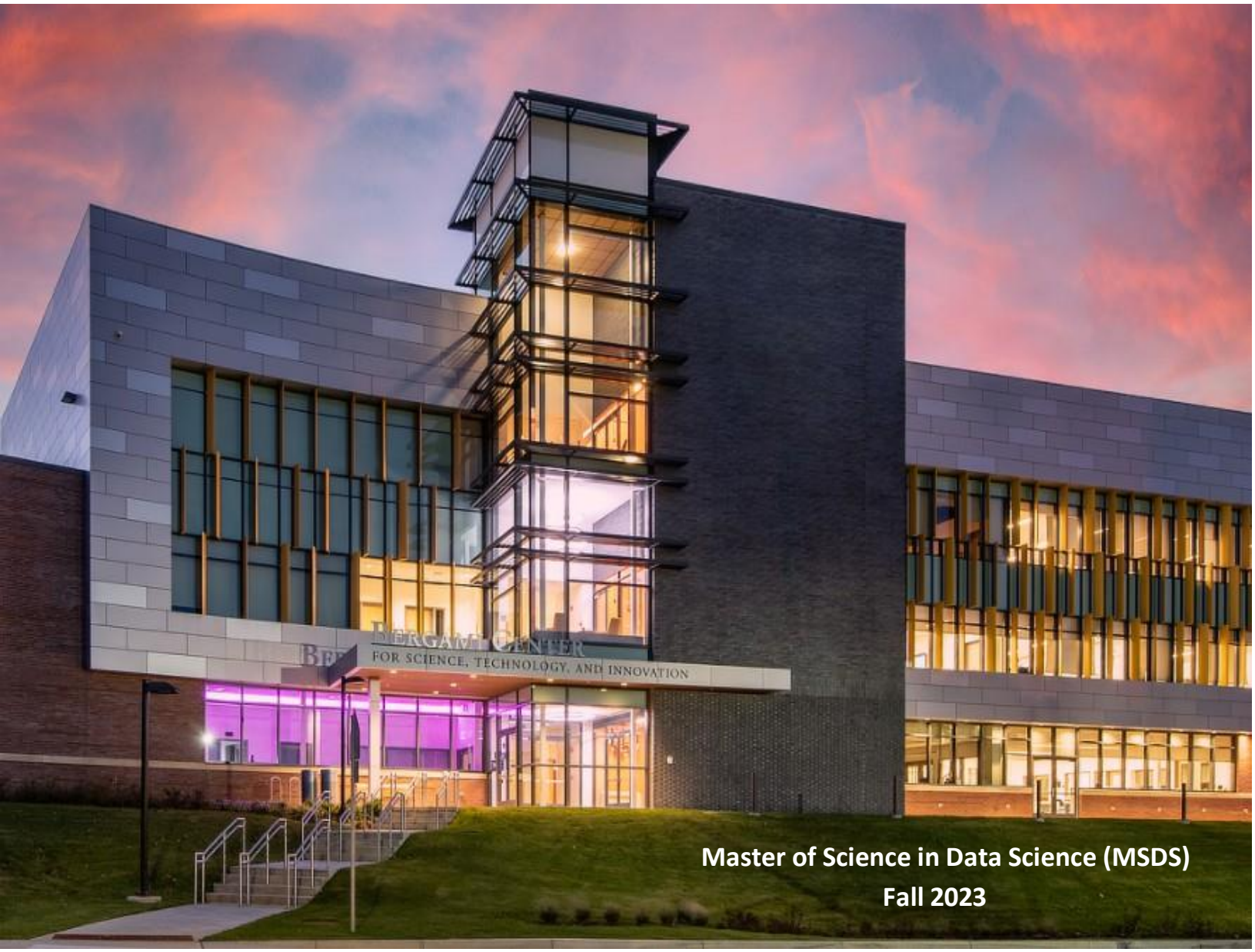
University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science



Master of Science in Data Science (MSDS)

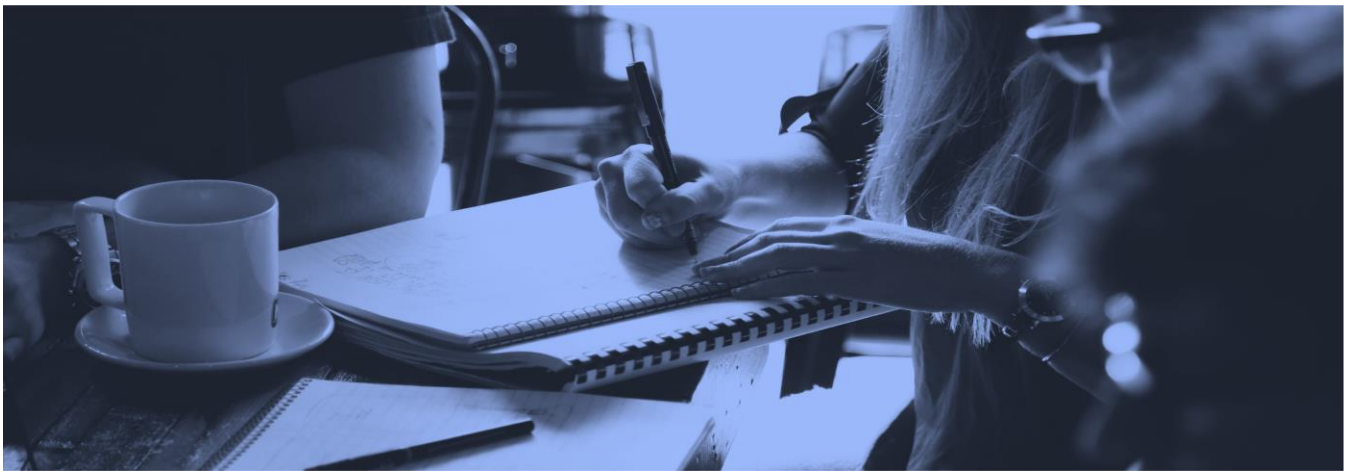
Fall 2023

CONTENTS

CROP PREDICTION

Executive Summary

Crop prediction's revolutionary effect on agriculture is highlighted. Crop prediction maximizes yields, improves resource efficiency, and reduces risk by utilizing technologies like machine learning and remote sensing. Farmers are able to make well-informed decisions that result in higher productivity, resource conservation, and lower costs by using data-driven insights. In overcoming obstacles and guaranteeing global food security, crop prediction plays a critical role in forming a robust and sustainable agricultural future.



Team Members:

HARI KRISHNA PARA
TURANGI SAI KRISHNA
RAHUL KUMAR MAVURI
SWETHA NUNEMUNTHALA

Questions?
Contact:
hpara2@unh.newhaven.edu
stura2@unh.newhaven.edu
rmavu1@unh.newhaven.edu
snune14@unh.newhaven.edu

CROP PREDICTION USING MACHINE LEARNING

Highlights of Project

Multi-Source Data Fusion: Integrating diverse datasets from sources like Kaggle, governmental repositories, IoT devices, and historical farm records to create a comprehensive data landscape for predictive modeling.

Predictive Model Development: Constructing robust machine learning models leveraging supervised learning techniques to forecast crop yields. These models consider factors like weather patterns, soil quality, and historical crop performance for accurate predictions.

Insightful Data Exploration: Conducting extensive exploratory data analysis (EDA) to unveil correlations, anomalies, and patterns within the agricultural data, enabling informed decision-making for farmers and stakeholders.

Ethical Data Handling: Ensuring data privacy and adhering to ethical standards when handling sensitive agricultural information, protecting farmer privacy while maximizing the potential of available data for predictive analytics.

Submitted on: 12-06-2023.

Abstract

Crop cultivation used to be done based on the practical experience of farmers. However, crop yields are already being negatively impacted by climate change. As a result, farmers are unable to select the best crop or crops depending on soil and environmental conditions, and the process of manually selecting the best crop or crops for a piece of land has typically failed. Crop production rises when crop predictions are accurate. This is where crop prediction is where machine learning comes into play. Crop forecasting is influenced by soil, geographical, and meteorological factors. Choosing suitable characteristics for the appropriate crop or crops is a fundamental component of the prediction process carried out by feature selection methods. An analysis of several wrapper feature selection techniques is done in this work.

Elevator Pitch Video Link:

Introductory Section:

The introduction of Crop Prediction represents a revolutionary step toward informed and optimized farming practices in the dynamic field of modern agriculture. By utilizing state-of-the-art technologies like IoT, remote sensing, and machine learning, Crop Prediction surpasses conventional approaches by offering farmers precise, data-driven insights into crop yields, resource allocation, and possible hazards. This introduction explores the revolutionary potential of Crop Prediction, highlighting how it can help farmers become more productive, adapt to the changing agricultural landscape, and adopt sustainable practices for a resilient and prosperous future.

Data Collection:

Gather historical data on crop yields, weather conditions, soil properties, and other relevant factors. Data sources can include agricultural surveys, remote sensing, meteorological databases, and on-field sensors.

Data processing:

1. Clean the data by handling missing values, outliers, and inconsistencies.
2. Convert data into a suitable format for analysis, ensuring that variables are appropriately scaled.

Exploratory Data Analysis:

performing exploratory analysis on the crop image dataset to understand the distribution of Temperature, CO2, Soil type, Soil EC, Rain fall, Humidity, Soil PH, NPK and other relevant characteristics.

Model Training and Evaluation:

Split the dataset into training and testing sets. Train the model on the training set, adjusting parameters as needed. Validate the model using the testing set to ensure it generalizes well to new data.
Performance Metrics and Assessing the model's performance on the validation set using the chosen metrics. This step helps in early detection of overfitting and guides further adjustments to the model

Methodology

Research Methods:

Research methods for crop prediction involve a combination of techniques from agricultural science, machine learning, and data science. Here is a comprehensive guide outlining the key research methods for crop prediction: we have used three popular machine learning algorithms they are logistic regression, random forest, decision tree. but random forest gave higher accuracy for prediction.

Random forest: Random Forest is a popular machine learning algorithm that falls under the category of ensemble learning methods. Ensemble learning involves combining the predictions of multiple models to improve overall performance and robustness. Random Forest, in particular, is known for its versatility and effectiveness in various types of prediction tasks, including regression and classification.

Data Sources:

We collected High-resolution image datasets from Kaggle, a well-known platform with a variety of datasets, are a major source of support for the project. Kaggle offers an extensive library of crop photos in a variety of settings, with different lighting and express. To enhance the Kaggle dataset, D lib, a toolkit that provides pre-trained models for crops and relevant images.

Data Collection Exercise:

Gather historical data on crop yields, weather conditions, soil properties, and other relevant factors. Data sources can include agricultural surveys, remote sensing, meteorological databases, and on-field sensors.

Choice for Variables, Data, and Methods:

Metadata variables: N, P, K, temperature, humidity, pH, rainfall, and crop yield (label). Data exploration: Assessing relationships, distributions, and correlations among variables. Understanding target variable: Analyzing crop yield patterns concerning input factors

Addressing Research Questions:

Research in crop prediction focuses on leveraging various factors to forecast crop yields accurately. This involves understanding how environmental aspects like weather conditions, soil quality, and pest/disease patterns influence crop growth. Researchers delve into massive datasets collected through remote sensing, IoT devices, and historical records to develop machine learning models that predict yields. These models aim to integrate multiple data sources and consider regional variations to offer insights crucial for farmers' decision-making. Additionally, exploring the impact of climate change on agriculture, optimizing crop rotation patterns for soil health, and developing user-friendly tools for farmers are key areas of focus. The ultimate goal is to create robust, adaptable models that assist farmers in maximizing yields sustainably while navigating the complexities of modern agriculture.

Studies also investigate economic aspects, such as market demand and profitability, to ensure these predictive models align with sustainable agricultural practices while benefiting the farming community. Collaborative efforts from experts in agriculture, data science, and environmental studies drive this research, aiming to provide accessible and reliable tools that aid farmers in making informed decisions for efficient and sustainable crop production.

CRISP-DM Methodology

Business Understanding:

Business Objective: "Optimize crop yield predictions for profitability."

Situation: Agricultural stakeholders in a specific region aim to enhance crop production while facing challenges posed by varying climate patterns, soil health issues, and market demands. They possess extensive historical data on weather, soil quality, crop types, and yields.

Data Science Goal: Develop a predictive model leveraging historical agricultural data to forecast crop yields accurately. This model should consider diverse factors such as weather patterns, soil health, and historical crop performance to assist farmers in making informed decisions regarding crop selection, rotation strategies, and resource allocation for maximizing yields and profitability in changing environmental conditions.

Data Understanding:

Data Sources:

The primary sources of data for this project are Kaggle. It provides a wealth of curated datasets relevant to crop prediction, offering access to diverse agricultural data such as weather patterns, soil compositions, historical crop yields, and related factors. This platform serves as a valuable resource for data scientists, enabling collaborative exploration, analysis, and the development of predictive models aimed at advancing agricultural practices and enhancing crop yield predictions.

Data Volume:
Kaggle datasets contain large volumes of structured and unstructured data, including historical weather records, soil quality assessments, crop-specific data, and market trends, potentially spanning terabytes of information.

Data Diversity:
Kaggle datasets are diverse, covering a range of data types, including numerical weather data, categorical crop information, satellite imagery (visual and spectral), textual soil reports, and market demand statistics, ensuring a comprehensive understanding of agricultural factors.

Data Labeling:
Kaggle Label historical data with crop types, yield quantities, disease occurrences, and pest infestations, facilitating supervised learning approaches for predictive modeling.

Data Privacy:
Ensure adherence to privacy regulations and ethical considerations when handling sensitive farmer information or proprietary agricultural data, employing anonymization techniques where necessary.

Data Exploration:
Conduct extensive exploratory data analysis (EDA) to understand correlations, patterns, and outliers within the datasets, employing visualization tools and statistical methods to gain insights into relationships among variables.

Data Imbalances:
Address potential imbalances in data distribution, particularly in cases where certain crops or regions might be underrepresented, employing techniques like oversampling, under sampling, or synthetic data generation to mitigate bias in predictive modeling.

Data Preparation:

Data cleaning: Handling missing values, outliers, and inconsistencies in the dataset. Feature engineering: Creating new features or transformations, if necessary, like aggregating weather data over specific periods. Data normalization or scaling: Ensuring variables are on a similar scale for modeling.

	N	P	K	temperature	humidity	ph	rainfall
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000
mean	50.551818	53.362727	48.149091	25.616244	71.481779	6.469480	103.463655
std	36.917334	32.985883	50.647931	5.063749	22.263812	0.773938	54.958389
min	0.000000	5.000000	5.000000	8.825675	14.258040	3.504752	20.211267
25%	21.000000	28.000000	20.000000	22.769375	60.261953	5.971693	64.551686
50%	37.000000	51.000000	32.000000	25.598693	80.473146	6.425045	94.867624
75%	84.250000	68.000000	49.000000	28.561654	89.948771	6.923643	124.267508
max	140.000000	145.000000	205.000000	43.675493	99.981876	9.935091	298.560117

Metadata variables: N, P, K, temperature, humidity, pH, rainfall, and crop yield (label).

Model Architecture:

Logistic Regression:

- Application: Predicting crop yield categories based on input variables.
- Model training: Utilizing logistic regression to fit the data.
- Parameter tuning: Adjusting model parameters for optimal performance.

```
LogReg = LogisticRegression()
LogReg.fit(x_train,y_train)

predicted = LogReg.predict(x_test)
x = metrics.accuracy_score(y_test,predicted)
acc.append(x)
model.append('Logistic Regression')
print("Logistic Regression Accuracy is",x * 100)
print(classification_report(y_test,predicted))
```

```
Logistic Regression Accuracy is 95.9090909090909
```

Random Forest:

- Application: Harnessing ensemble learning for improved yield prediction.
- Model development: Building multiple decision trees and aggregating predictions.
- Parameter optimization: Tuning hyperparameters for enhanced accuracy.

```
RF = RandomForestClassifier(n_estimators=29, criterion = 'entropy',random_state=0)
RF.fit(x_train,y_train)
predicted = RF.predict(x_test)
x = metrics.accuracy_score(y_test,predicted)
acc.append(x)
model.append('Random Forest')
print("Random Forest Accuracy is ",x * 100)
print(classification_report(y_test,predicted))
```

```
Random Forest Accuracy is  99.0909090909091
```

Decision Trees:

- Application: Employing decision trees for straightforward interpretability in yield prediction.
- Model creation: Constructing decision tree-based models using the metadata.
- Interpretability: Analyzing the decision-making process of the trees.

```
dt_model.fit(x_train.values,y_train.values)
tree_predicted = dt_model.predict(x_test)

# Calculate accuracy
tree_accuracy = metrics.accuracy_score(y_test, tree_predicted)
print("Decision Tree Accuracy is", tree_accuracy * 100)
```

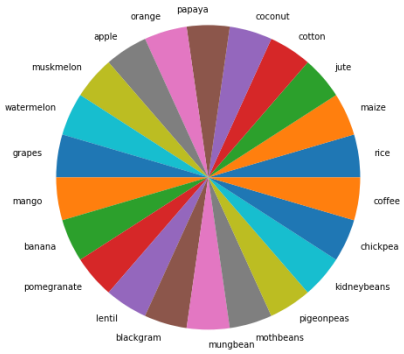
```
Decision Tree Accuracy is 98.86363636363636
```

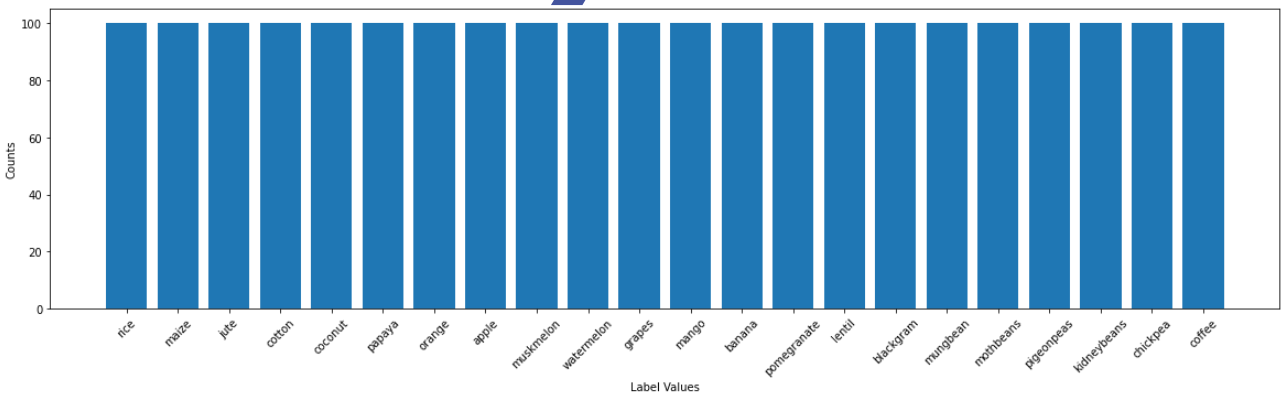
Results:

Model Evaluation and Results:

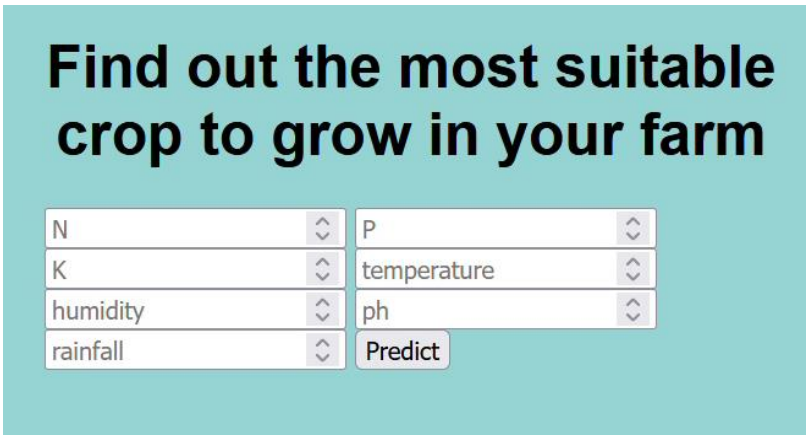
We are evaluating the trained model on the validation set and presenting performance metrics such as accuracy, precision, recall, and the confusion matrix. Visualization of the confusion matrix provides insights into how well the model is performing across different classes.

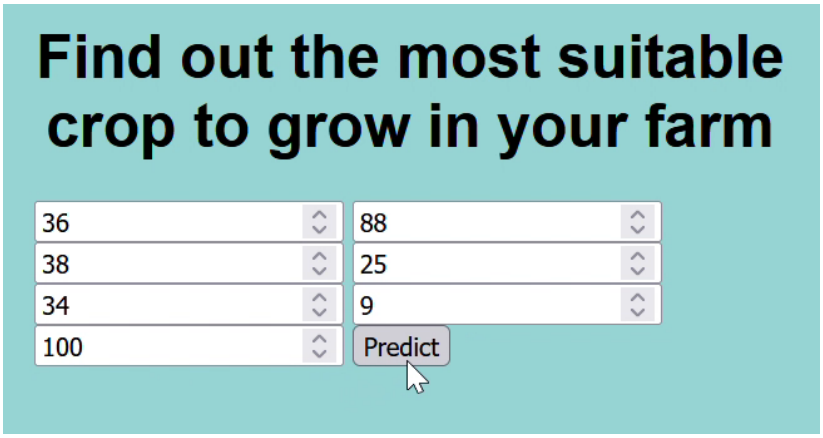
Visualization of the Model:

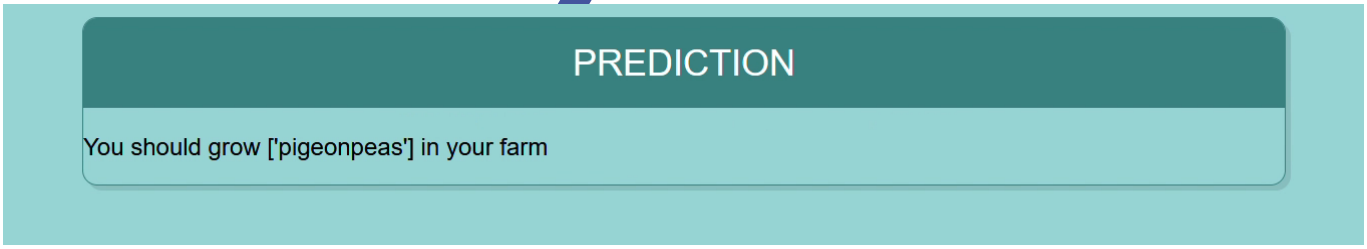




Model Deployment and Prediction Example:







- Model selection: Choosing the most suitable algorithm based on evaluation metrics and requirements.
- Final model fine-tuning: Optimizing the selected model for the best performance.
- Implementation: Integrating the selected model into a user-friendly platform for stakeholders

Conclusion

Crop prediction offers a strategic advantage in agriculture by leveraging data-driven insights It facilitates efficient resource allocation, aiding in optimal water, fertilizer, and pesticide usage. Predictive models assist in mitigating risks posed by unpredictable weather, pests, and diseases. Accurate yield forecasts enable informed financial planning, guiding investments and pricing strategies
Embracing crop prediction supports sustainable agriculture, reducing waste and environmental impact. This approach contributes significantly to global food security by estimating food production and distribution. Ongoing innovation and the integration of advanced technologies will shape a more resilient future for agriculture

Contributions/References

Gehlot, A.; Sidana, N.; Jawale, D.; Jain, N.; Singh, B.P.; Singh, B. Technical analysis of crop production prediction using Machine Learning and Deep Learning Algorithms. In Proceedings of the International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 24–25 September 2022; pp. 1–5. [Google Scholar]
Vashisht, S.; Kumar, P.; Trivedi, M.C. Improvised Extreme Learning Machine for Crop Yield Prediction. In Proceedings of the 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, UK, 27–29 April 2022; pp. 754–757. [Google Scholar]
Dean, J. The deep learning revolution and its implications for computer architecture and chip design. In Proceedings of the IEEE International Solid-State Circuits Conference-(ISSCC), San Francisco, CA, USA, 16–20 February 2020. [Google Scholar]
Shahin, F.; Zahin, L.; Rahman, R.; Hossain, A.J.; Kaf, A.H.; Abdul Malek Azad, A.K.M. Agricultural Analysis and Crop Yield Prediction of Habiganj using Multispectral Bands of Satellite Imagery with Machine Learning. In Proceedings of the 11th International Conference on Electrical and Computer Engineering (ICECE), Dhaka, Bangladesh, 17–19 December 2020; pp. 21–24. [Google Scholar]
Tawseef, A.S.; Tabasum, R.; Faisal, R.L. Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. Computer. Electron. Agric. 2022, 198, 107119. [Google Scholar]
Vivek, S.; Ashish, K.T.; Himanshu, M. Technological revolutions in smart farming: Current trends, challenges & future directions. Computer Electron. Agric. 2022, 201, 107217. [Google Scholar]
Mamatha, J.C.K. Machine learning based crop growth management in greenhouse environment using hydroponics farming techniques. Meas. Sens. 2023, 25, 100665.