Vol.1, Issue.2, (2024)
Journal of Techno Trainers
https://www.technotrainers.net

Crop recommendation system using machine learning. a data-driven approach to sustainable agriculture

Syed Bilawal Bukhari[1]
[1][Department of Artificial intelligence, The Islamia university of Bahawalpur]

Article Information

Article Type: Research Article

ABSTRACT

In recent years, the agricultural sector has witnessed a significant transformation with the infusion of advanced technology and analytics, aiming to optimize farming processes and yield. This research dives deep into the possibilities of leveraging machine learning techniques to recommend suitable crops based on environmental factors. Using a dataset enriched with multiple parameters such as temperature, humidity, rainfall, and soil pH, a predictive model was constructed. The cornerstone of our study was the implementation of the Random Forest Classifier, owing to its robustness in handling complex datasets and delivering reliable predictions. Experimentation was thorough, considering various preprocessing techniques like Minmax scaling and standard normalization to refine our data for the model. Additionally, a comparative analysis was conducted across multiple machine-learning algorithms to ensure the efficacy of the chosen method. This crop recommendation system stands as a testament to the powerful synergy of agriculture and technology, promising to pave the way for future data-driven innovations in the field. By ensuring that the crops selected are in harmony with their environmental conditions, this system not only bolsters agricultural productivity but also paves the path for sustainable farming.

Keywords:

Crop Recommendation System, Machine Learning, Random Forest Classifier, Environmental Parameters in Crop Selection, Predictive Modeling in Farming

INTRODUCTION

Agriculture, as an age-old practice, has been the backbone of numerous economies and has provided sustenance to billions (Falkeis, Shesterikova, James, & Tingen, 2022). However, as global challenges like population growth, climate change, and diminishing resources emerge, it's evident that traditional farming methods might not suffice to ensure food security in the future (Muluneh, 2021). The present-day scenario mandates a more analytical and informed approach towards farming(Bryant et al., 2020). It's here that the confluence of agriculture and technology, particularly artificial intelligence and machine learning, offers promising solutions. With the exponential growth of data in the agriculture domain, ranging from satellite imagery to sensors that provide real-time soil and weather data, there's a pressing need to process this data intelligently(Amudha, 2021). Traditional data analysis methods fall short in handling this magnitude and complexity. Machine learning, with its ability to unearth patterns and insights from vast datasets, emerges as the linchpin in this endeavor(Jayapandian, 2023). Artificial Intelligence (AI) and Machine Learning (ML) offer unprecedented tools in this endeavor. These

technologies introduce an amalgamation of tradition and modernity into agriculture(Júnior et al., 2024). This blend's significance lies in its potential to elevate yield, profitability, and promote sustainable farming practices. By tailoring recommendations specific to an agricultural area or farm's conditions, we can ensure that the advice is both personalized and effective. (Nag, Das, Chand, & Roy, 2024). With the advent of technology and the increasing demand for food due to the growing global population, there is a pressing need to optimize agricultural practices. Traditional farming relies on human judgment, often based on years of experience and intuition, to make decisions regarding which crops to plant under varying environmental conditions(Jhariya, Meena, & Banerjee, 2021). While this method has been effective for centuries, the changing climatic patterns and the sheer unpredictability of the environment make these decisions more challenging. This dissertation introduces a detailed and comprehensive crop recommendation system that caters to various parameters, including soil health, climate conditions, and economic considerations.

Methodology:

The methodology chapter will provide a detailed description of the research design, data collection, data analysis, and the methods used to conduct the study. The steps and procedures used to obtain the results will be outlined to give readers a comprehensive understanding of the study's design.**Research Design**

At the heart of our research lies, a quantitative design anchored in data-driven methodology. Our dataset, meticulously assembled from a spectrum of trusted sources, provides a holistic picture of agricultural dynamics. This information spans a wide array of attributes from volatile weather patterns to complex soil quality parameters, and other critical agronomic variables pivotal to crop health and yield.

Our methodology seamlessly bridges the rich legacy of traditional farming wisdom with the avant-garde capabilities of contemporary technology. By integrating these two realms, we envision a transformative shift in agriculture. Farmers, instead of solely relying on instinct or ancestral practices, will have the arsenal of empirical data at their disposal, marking the dawn of precision agriculture.

In essence, our research design, fortified with a clear methodological framework and promising outcomes, strives to be a catalyst in the metamorphosis of contemporary agricultural practices.

Data Collection

Data Collection: The crux of any data-driven study is the quality and relevance of the dataset used. For our research, we've leveraged a comprehensive dataset aggregated from myriad agricultural sources. This dataset offers a detailed record of diverse crop yields subjected to a range of environmental and soil conditions.

Attributes of the Dataset: The dataset comprises several pivotal attributes, each playing a crucial role in determining the viability and yield of specific crops:

- **Nitrogen, Phosphorus, and Potassium**: Often collectively referred to as the primary macronutrients, these elements play vital roles in plant nutrition. Their quantities can significantly affect crop yield and quality.

- **Temperature**: A significant environmental factor, temperature variations can impact the growth cycle of crops. Certain crops thrive in warm temperatures, while others prefer cooler climes.
- **Humidity**: Humidity levels can influence the rate of crop transpiration and photosynthesis. It's an integral parameter, especially for moisture-sensitive crops.
- **pH Level**: The pH level of the soil can affect the solubility of minerals and nutrients. Crops tend to flourish best at specific pH ranges.
- **Rainfall**: The amount and distribution of rainfall can dramatically impact crop health and yield. While certain crops require substantial water, others can thrive in drier conditions.

Preprocessing and Statistical Operations

Before deploying our primary machine learning model, we embarked on an essential journey of data preprocessing and statistical analysis. Understanding and preparing our data was paramount to ensure the robustness of our subsequent analyses. The following are the key steps and operations we undertook:

1. **Data Cleaning:** We began by identifying and handling any missing values, outliers, or anomalies present in our dataset. This ensured that our dataset was free from any discrepancies that could hinder the performance of our model.
2. **Exploratory Data Analysis (EDA):** Using various statistical tools and visualization techniques, we conducted an EDA to garner insights from the data. This step allowed us to understand the distribution of variables, detect patterns, and identify any potential correlations or relationships.
3. **Feature Engineering:** Based on our EDA findings, we engaged in feature engineering, where we selected, transformed, or even created new variables that could enhance the predictive power of our model.
4. **Normalization/Standardization:** Given the diverse range of variables in our dataset, we applied normalization or standardization techniques to ensure all variables were on a comparable scale. This is essential for algorithms, especially those that rely on distance metrics, like K-Nearest Neighbors.

```
array([[-9.03426596e-01, -1.12616170e+00, -6.68506601e-01, ...,
         9.36586183e-01,  1.93473784e-01,  5.14970176e-03],
       [-3.67051340e-01,  7.70358846e-01, -5.70589522e-01, ...,
        -1.00470485e-01,  8.63917548e-01, -6.05290566e-01],
       [-1.17161422e+00,  5.89737842e-01, -4.53089028e-01, ...,
        -3.82774991e-01,  1.05029771e+00, -1.04580687e+00],
       ...,
       [-1.06433917e+00, -5.24091685e-01, -3.35588533e-01, ...,
        -8.98381379e-01, -6.34357580e-04, -4.37358211e-02],
       [-1.06433917e+00,  2.12501638e+00,  3.05234239e+00, ...,
         3.86340190e-01, -1.48467347e-01, -5.69036842e-01],
       [-5.01145154e-01,  7.40255346e-01, -5.11839275e-01, ...,
        -4.18045489e-01,  6.86860180e-01, -8.96531475e-01]])
```

5. **Statistical Tests:** To validate our findings and to understand the significance of various attributes, we executed a series of statistical tests. These tests provided a robust backing to our initial observations and findings from the EDA.

6. **Data Splitting:** Post all preprocessing steps, we divided our dataset into training and testing subsets. This ensured that our model was evaluated on unseen data, providing a genuine measure of its performance.

By investing considerable effort in these preliminary stages, we laid a solid foundation for our primary research activities. Such rigorous data preparation not only bolstered the reliability of our subsequent analyses but also reinforced the credibility of our research outcomes.

	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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2200 entries, 0 to 2199
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   N            2200 non-null   int64
 1   P            2200 non-null   int64
 2   K            2200 non-null   int64
 3   temperature  2200 non-null   float64
 4   humidity     2200 non-null   float64
 5   ph           2200 non-null   float64
 6   rainfall     2200 non-null   float64
 7   label        2200 non-null   object
dtypes: float64(4), int64(3), object(1)
memory usage: 137.6+ KB
```

```
crop.isnull().sum()
✓ 0.0s

N              0
P              0
K              0
temperature    0
humidity       0
ph             0
rainfall       0
label          0
dtype: int64
```

Model Training in Our Study

In our specific study, model training was approached with precision, ensuring that every decision and step was tailored to our dataset and research objectives. Here's an in-depth breakdown:

- **Data Pre-processing**: One of the primary steps before delving into model training was data preprocessing. We normalized and standardized our data to ensure uniformity in the dataset. This was achieved using **Min Max Scaler** for normalization and **Standard Scaler** for standardization, making sure that no particular feature dominated the others in terms of scale.

- **Model Selection**: Based on the nature of our dataset and the problem at hand, we employed the Random Forest Classifier. The choice was influenced by its robustness against overfitting, its capability to handle large datasets, and its prowess in handling both classification and regression tasks effectively.

- **Model Training**: The preprocessed dataset was then split into training (**X_train** and **y_train**) and testing sets. The training set was fed into the Random Forest Classifier, adjusting its numerous decision trees to best predict the target variable. The algorithm seeks the best feature among a random subset of features, splitting the node from the best result, which in turn yields a forest of trees.

- **Model Evaluation**: After training, the model's effectiveness was gauged using the test set (**X_test**). This step was vital to understand the model's ability to generalize to unseen data. The main metric used for this evaluation was accuracy, highlighting the percentage of correct predictions made by the model.

```
X_train
✓ 0.0s
array([[0.12142857, 0.07857143, 0.045     , ..., 0.9089898 , 0.48532225,
        0.29685161],
       [0.26428571, 0.52857143, 0.07      , ..., 0.64257946, 0.56594073,
        0.17630752],
       [0.05      , 0.48571429, 0.1       , ..., 0.57005802, 0.58835229,
        0.08931844],
       ...,
       [0.07857143, 0.22142857, 0.13      , ..., 0.43760347, 0.46198144,
        0.28719815],
       [0.07857143, 0.85      , 0.995     , ..., 0.76763665, 0.44420505,
        0.18346657],
       [0.22857143, 0.52142857, 0.085     , ..., 0.56099735, 0.54465022,
        0.11879596]])
```

Accuracy in Model Evaluation

In the realm of machine learning, the accuracy of a model is a vital metric, especially when dealing with classification tasks like ours. It quantifies how often the model's predictions align with the actual outcomes.

Defining Accuracy: Accuracy is defined as the ratio of the number of correct predictions made by the model to the total number of predictions. Mathematically, this can be represented by the equation:

$$\boldsymbol{Accuracy = \frac{Total\ Number\ of\ Predictions}{Number\ of\ Correct\ Predictions}}$$

In the context of our study:

- **Number of Correct Predictions**: These are instances where our model's prediction for a specific data point (like the ideal crop type for given conditions) perfectly matched the actual outcome in our test set.
- **Total Number of Predictions**: This denotes the complete set of predictions made by our model on the test dataset.

Significance in our Study: The accuracy metric played a pivotal role in our research. By evaluating the accuracy post-training, we could deduce the Random Forest Classifier's capability to predict the best crop type under diverse conditions. A high accuracy indicates that our model is well-calibrated with the training data and has managed to capture the underlying patterns, hence is more likely to make accurate predictions on new, unseen data.

```
Logistic Regression  with accuracy : 0.9636363636363636
Naive Bayes  with accuracy : 0.9954545454545455
Support Vector Machine  with accuracy : 0.9681818181818181
K-Nearest Neighbors  with accuracy : 0.9590909090909091
Decision Tree  with accuracy : 0.9886363636363636
Random Forest  with accuracy : 0.9954545454545455
Bagging  with accuracy : 0.9886363636363636
AdaBoost  with accuracy : 0.1409090909090909
Gradient Boosting  with accuracy : 0.9818181818181818
Extra Trees  with accuracy : 0.8931818181818182
```

Deployment

After the model training, validation, and finalization phases, the next essential step in our research methodology was deployment. Deploying the trained model is critical as it enables the application to be used in real-world scenarios, offering actionable insights to the end-users.

In our case, the trained model was encapsulated into a web-based application. This application acts as an interface between the user and our predictive system. Given the ubiquity and accessibility of web platforms, it was deemed the most effective way to reach a broad audience and ensure seamless interaction.

User Interface:

The primary interaction point of our deployment is a user-friendly web page. This page is designed with simplicity and effectiveness in mind. It comprises a form where users can input specific parameters related to agricultural conditions:

- **Nitrogen**
- **Phosphorus**
- **Potassium**
- **Temperature**
- **Humidity**
- **pH Level**
- **Rainfall**

Upon entering these parameters, users can initiate the prediction process. The backend system, leveraging our trained model, processes these inputs and returns the most suitable crop prediction based on the entered conditions. The predicted crop is then dynamically displayed on the webpage, offering users immediate insights into their agricultural decisions.

The inclusion of a web-based interface not only simplifies the prediction process but also makes the technology accessible to users regardless of their technological proficiency. Through this system, farmers, agricultural experts, and enthusiasts can make informed decisions, ensuring optimal yields and efficient farming practices.

Ethical Considerations

In the execution and application of our predictive crop cultivation system, various ethical facets come to the forefront:

- **Data Integrity and Confidentiality**: The agricultural datasets employed in our study have been responsibly sourced, ensuring there is no inclusion of personally identifiable or sensitive information. Any potential identifiers within the data have undergone anonymization processes, upholding the essence of data confidentiality.
- **Transparent Methodology**: All methodologies, techniques, and results within this research have been documented straightforwardly. Our commitment is to provide a clear and honest depiction of our work, inclusive of both its strengths and limitations.
- **Bias Mitigation**: A core focus throughout this research has been to ensure that our predictive model is devoid of biases. We've taken measures to ensure our training data is representative, and our prediction algorithms are continually assessed for any unintended biases.
- **Impact Awareness**: Recognizing the potential economic ramifications of our predictions for farmers and other agricultural stakeholders, we've endeavored for our system to be as precise as possible, ensuring it's corroborated with real-world data and expert insights.
- **Feedback and Iteration**: Post the web page deployment for our prediction system, users have the provision to offer feedback. This feedback loop ensures that the system is regularly updated, and any inconsistencies are rectified promptly.
- **Ongoing Ethical Adherence**: Even post-deployment, we commit to an ongoing review of our system's societal impact, assuring it remains in line with the predefined ethical standards and adjusting its course as and when required.

Ensuring that our predictive system is not just technically sound but also ethically robust has been of paramount importance throughout this research. Our objective is to provide an efficient tool while upholding the highest ethical standards.

Results and discussion:

Recapitulation

The core objective of our study was to assist farmers in making informed decisions about crop cultivation based on specific environmental parameters. Our journey began with the meticulous gathering of a diverse dataset, followed by the construction of a predictive model, and culminated in the creation of a user-friendly web application.

Major Achievements

- Data Collection and Preprocessing: We amassed a significant dataset from various agricultural sources, emphasizing its thoroughness and accuracy. The preprocessing phase eliminated anomalies and ensured a streamlined data flow, ready for subsequent stages.

- Modeling and Validation: Our model's efficiency was accentuated by its robustness in predicting the most optimal crops. It underwent several rounds of validation to ensure its predictions matched real-world expectations.

- Web Deployment: The launch of our interactive web application marked a significant milestone. It made complex predictive algorithms accessible to farmers, translating high-end computations into actionable insights.

Lessons Learned

The entire journey underscored several lessons. First and foremost, the importance of data integrity; the saying, "garbage in, garbage out" became more profound. Another pivotal lesson was the interdisciplinary nature of applied research; bridging data science with agriculture necessitated a holistic approach, demanding inputs from agronomists, data scientists, and software developers.

Implications for Stakeholders

The potential of our tool extends beyond immediate agricultural practices. Policy-makers can leverage the insights garnered to frame region-specific agricultural policies. Agri-businesses can align their products and services based on predicted crop trends. Additionally, educational institutions can utilize our tool as a practical case study, bridging theoretical knowledge with real-world applications(Gardezi et al., 2024).

Environmental Considerations

The planet's changing climate necessitates sustainable agricultural practices. By recommending crops that are best suited to prevailing environmental conditions, we reduce the strain on resources like water and soil. Over time, aligning crop cultivation with environmental parameters can contribute to sustainable farming, reducing the carbon footprint and ensuring food security in the face of climatic uncertainties.

Limitations and Challenges

Every research project faces its set of challenges, and ours was no exception. Data collection, while extensive, might not capture the entirety of the diverse agricultural landscape. Additionally, while our model predictions are accurate, there are always external variables, such as sudden climatic events or pest outbreaks, that can influence outcomes. These limitations underline the importance of continuous model updates and real-world validations.

Conclusion:

In an era marked by rapid technological advancements, it's imperative to ensure that the benefits permeate all sectors, agriculture included. Our research is a modest step in that direction. While we've made significant headway, the journey is ongoing. Agriculture's symbiotic relationship with nature means it's ever-changing, and our tools need to evolve in tandem to remain relevant.

Author contributions: All authors equally contributed to this study
Competing Interests: The author declares that this work has no competing interests.
Grant/Funding information: The author declared that no grants supported this work.
Data Availability Statement: The associated data is available upon request from the corresponding author.

REFERENCES

Amudha, T. (2021). Artificial intelligence: a complete insight. In *Artificial intelligence theory, models, and applications* (pp. 1-24): Auerbach Publications.

Bryant, B. P., Kelsey, T. R., Vogl, A. L., Wolny, S. A., MacEwan, D., Selmants, P. C., . . . Butterfield, H. S. (2020). Shaping land use change and ecosystem restoration in a water-stressed agricultural landscape to achieve multiple benefits. *Frontiers in Sustainable Food Systems, 4*, 138.

Falkeis, A., Shesterikova, A., James, B., & Tingen, M. (2022). *Nonlinear Urbanism: Towards Multiple Urban Futures*: Birkhäuser.

Gardezi, M., Abuayyash, H., Adler, P. R., Alvez, J. P., Anjum, R., Badireddy, A. R., . . . Dadkhah, A. (2024). The role of living labs in cultivating inclusive and responsible innovation in precision agriculture. *Agricultural Systems, 216*, 103908.

Jayapandian, N. (2023). *Machine Learning Based Spam E-Mail Detection Using Logistic Regression Algorithm.* Paper presented at the 2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG).

Jhariya, M. K., Meena, R. S., & Banerjee, A. (2021). Ecological intensification of natural resources towards sustainable productive system. *Ecological intensification of natural resources for sustainable agriculture*, 1-28.

Júnior, M. R. B., de Almeida Moreira, B. R., dos Santos Carreira, V., de Brito Filho, A. L., Trentin, C., de Souza, F. L. P., . . . Ampatzidis, Y. (2024). Precision agriculture in the United States: A comprehensive meta-review inspiring further research, innovation, and adoption. *Computers and Electronics in Agriculture, 221*, 108993.

Muluneh, M. G. (2021). Impact of climate change on biodiversity and food security: a global perspective—a review article. *Agriculture & Food Security, 10*(1), 1-25.

Nag, A., Das, A., Chand, N., & Roy, N. (2024). Sustainable Agriculture: A Critical Analysis of Internet of Things—Based Solutions. In *Intelligent Systems and Industrial Internet of Things for Sustainable Development* (pp. 118-138): Chapman and Hall/CRC.