Crop Yield Prediction in India

Rishabh Awasthy - MS. Data Science '23
Mentor: Dr. Christelle Scharff
Pace University, Seidenberg School of CSIS

Python Notebook : - Link
Dataset : - Link
GitHub Link

Abstract

Crop yield prediction is a critical challenge in modern agriculture, given its profound implications for global food security and economic stability. The ability to accurately forecast yields enables farmers, policymakers, and stakeholders to make informed decisions, optimize resource allocation, and mitigate the impacts of climate change on agriculture. Despite the significance of this problem, existing research exhibits a notable gap in achieving precise predictions, primarily due to limitations in incorporating diverse and dynamic factors affecting crop growth. This study addresses the identified gap by proposing a novel approach to crop yield prediction that integrates advanced machine learning techniques with a comprehensive set of environmental, agronomic, and historical data.
To validate the effectiveness of the proposed solution, extensive evaluations are conducted using a diverse range of datasets from different geographical locations..
The primary contribution of this research lies in presenting a novel and effective solution to the persistent challenge of crop yield prediction. By bridging the existing gap in related work, this study not only advances the field of agricultural data science but also provides a practical tool for stakeholders to enhance decision-making processes, optimize resource utilization, and ultimately contribute to global food security.

Research Question

Can we predict crop yield per acre, using data on farming practices and environmental conditions?

Dataset

The dataset is collected by Digital Green from smallholder farmers in India, focuses on addressing the challenges faced by these farmers, who play a vital role in global food production but often grapple with poverty and malnutrition. Comprising **44 columns and 3870 rows**, the dataset includes essential agricultural information such as **land characteristics of different areas, farming practices,** and **environmental conditions.** Noteworthy features encompass **land preparation methods, irrigation details, fertilizer application, and post-harvest practices**. The **'Acre'** and **'Yield'** columns represent the land size in acres and the corresponding crop yield per acre, respectively.

Methodology

This code initiates a comprehensive data analysis pipeline, encompassing profiling, correlation visualization, feature selection, and machine learning model building for predicting crop yield, fostering informed decision-making in agriculture.

1.Data Profiling:

1. Utilized the pandas_profiling library to generate a comprehensive report on the dataset (data.profile_report()).
2. Enabled sorting, full-width HTML style, and a progress bar for better visualization and understanding.

2.Correlation Analysis:

1. Employed Seaborn and Matplotlib for visualizing the correlation matrix using a heatmap.
2. Identified features correlated with the target variable ('Yield') to understand feature importance.

3.Feature Selection Based on Correlation:

1. Set a correlation threshold (0.20) to filter features significantly correlated with 'Yield.'
2. Selected features meeting or exceeding the correlation threshold for further analysis and model building.

4.Handling Missing Data:

1. Investigated missing data, identifying columns with null values.
2. Imputed missing values using a strategy of replacing nulls with correlated features, enhancing data completeness.

5.Replacement Strategy for Null Values:

1. Implemented a targeted approach for replacing nulls, ensuring data integrity.
2. Focused on replacing missing values in specific columns ('2tdUrea', '1tdUrea', BasalDAP, BasalUrea) by comparing correlations among these features.

6.Model Building: - Diverse machine learning models, including deep learning with TensorFlow and PyTorch, as well as traditional models like KNN, ElasticNet, Random Forest, Decision Tree, Support Vector Regressor, and Linear Regression, were utilized for predicting crop yield.

Results

Model evaluation highlights varied predictive accuracies in crop yield estimation.
•Decision Tree: Moderate accuracy (MAE: 158.63).
•DNN using TensorFlow: Improved accuracy (MAE: 144.75), showcasing pattern capture.
•KNN with PyTorch: Competitive accuracy (MAE: 142.15), outperforming traditional KNN.
•Traditional KNN: Higher MAE (245.64) than its PyTorch-based counterpart.
•Elastic Net: Moderate accuracy (MAE: 214.26), between Decision Tree and KNN.
•Random Forest: Enhanced accuracy (MAE: 141.68), demonstrating robustness.
•SVR: Higher MAE (274.75), indicating challenges in pattern extraction.
•Linear Regression: Moderate accuracy (MAE: 173.46).

In summary, the Random Forest and KNN using PyTorch models demonstrated relatively lower mean absolute errors, suggesting their efficacy in predicting crop yield in comparison to other algorithms. The results highlight the importance of selecting appropriate models for the specific characteristics of the dataset.

Model	Mean Absolute Error (MAE)
Decision Tree	158.63
DNN Using TensorFlow	144.75
KNN Using PyTorch	142.15
KNN	245.64
Elastic Net Model	214.26
Random Forest	141.68
SVR (Support Vector Regressor)	274.75
Linear Regression	173.46

Conclusion, Future Work & Limitation

Conclusions: The study successfully addressed the research question of predicting crop yield per acre, showcasing varying accuracies among machine learning models. Notably, Random Forest and KNN with PyTorch demonstrated superior performance, emphasizing the importance of model selection. Decision Tree, DNN using TensorFlow, and Elastic Net provided valuable insights, while Linear Regression served as a baseline. These findings underscore the need for tailored model choices aligned with dataset characteristics.

Future Work: Future research avenues involve exploring feature engineering, incorporating temporal analysis for evolving patterns, utilizing ensemble methods, and optimizing hyperparameters. Additionally, the inclusion of domain-specific features could enhance predictive capabilities.

Limitations: Limitations include data constraints, potential non-linearity not captured by linear models, and the need for validation on diverse datasets and regions. Incomplete feature sets and external factors, such as unforeseen events, may impact predictions. Addressing these limitations and pursuing future avenues will contribute to advancing accurate crop yield predictions, supporting informed decision-making in agriculture.

References

- S. M. M. Nejad, D. Abbasi-Moghadam, A. Sharifi, N. Farmonov, K. Amankulova and M. Lászlź, "Multispectral Crop Yield Prediction Using 3D-Convolutional Neural Networks and Attention Convolutional LSTM Approaches," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 16, pp. 254-266, 2023, doi: 10.1109/JSTARS.2022.3223423.
- P. Saini and B. Nagpal, "Deep-LSTM Model for Wheat Crop Yield Prediction in India," *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, Sonepat, India, 2022, pp. 73-78, doi: 10.1109/CCiCT56684.2022.00025.
- R. Welekar and C. Dadiyala, "Optimizing Crop Yield in Agriculture using Data Mining and Machine Learning Techniques," *2023 4th International Conference for Emerging Technology (INCET)*, Belgaum, India, 2023, pp. 1-7, doi: 10.1109/INCET57972.2023.10170493.
- S. Thirumal and R. Latha, "Automated Rice Crop Yield Prediction using Sine Cosine Algorithm with Weighted Regularized Extreme Learning Machine," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 35-40, doi: 10.1109/ICICCS56967.2023.10142403.