ENHANCING CROP YIELD PREDICTION THROUGH MACHINE LEARNING TECHNIQUES

Project Reference No.: *47S_MSC_0133*

College ***:*** *N.M.K.R.V College for Women, Bengaluru*
Branch ***:*** *Department of Data Science*
Guide(s) ***:*** *Prof. Deepalakshmi R.*
Student(S) ***:*** *Ms. Kauser Nissar T. P.*
Ms. Ashwitha M. R.
Ms. Shaziya Sultana

Abstract

Crop yield prediction is crucial for agricultural planning, impacting food security, economic stability, and resource optimization. Leveraging machine learning (ML) enhances the accuracy and efficiency of predicting crop outputs based on various factors. Accurate predictions are vital for adaptive strategies in the face of climate change, optimizing resource usage, and ensuring sustainable agriculture. This study proposes the use of a Temporal Fusion Transformer (TFT) model, designed specifically for time series forecasting, to enhance the accuracy and interpretability of crop yield predictions. The methodology involves comprehensive data collection, model development, feature selection, and optimization, followed by model validation and deployment. The proposed approach offers a promising solution for improving crop yield predictions and supporting informed decision-making in agriculture.

Keywords

Crop yield prediction, machine learning, Temporal Fusion Transformer (TFT), time series forecasting, feature engineering, model validation, sustainable agriculture.

Introduction

Crop yield prediction is essential for agricultural planning, impacting food security, economic stability, and resource optimization. Leveraging machine learning (ML) enhances the accuracy and efficiency of predicting crop outputs based on various factors[1]. Accurate crop yield predictions are crucial for resource planning and distribution to meet the growing food demands of a population expected to reach nearly 10 billion by 2050[3]. ML models analyze historical climate data to predict yields under different climate scenarios, aiding adaptive strategies for climate change. Informed decisions on planting, crop selection, and investments reduce the risks of overproduction or underproduction, stabilizing market prices[4]. ML predicts optimal usage of water, fertilizers, and pesticides, minimizing waste and costs. ML integration enables site-specific crop management, improving yield quality and quantity by analyzing data from sources like satellite imagery and IoT devices[2]. Additionally, ML algorithms identify patterns in vast datasets, leading to new discoveries in crop behaviour and environmental responses. Optimized input usage through ML reduces runoff and pollution, promoting sustainable farming, while accurate predictions improve land-use planning, preserving habitats and promoting biodiversity[4]. ML models also predict crop failures or yield reductions due to pests, diseases, or adverse weather, enabling proactive risk mitigation, and support better insurance products for farmers, ensuring financial stability[3]. Studying crop yield prediction using ML is vital for modern agriculture, offering tools to optimize production, ensure food security, and promote economic and environmental sustainability amid growing challenges from population pressures and climate change[2].

To enhance the accuracy and interpretability of crop yield predictions, we propose using a Temporal Fusion Transformer (TFT). The TFT is a state-of-the-art model specifically designed for time series forecasting, combining recurrent layers for learning local dependencies and self-attention mechanisms for capturing long-term dependencies and interactions within the data. This approach is particularly suitable for handling the complex relationships inherent in agricultural data. This methodology addresses the limitations of traditional models, offering improved predictive accuracy and interpretability, which are critical for informed decision-making in agriculture. [5]

Related Works

Crop yield prediction remains a complex challenge in agriculture, influenced by diverse factors such as soil quality, temperature, humidity, seed quality, and rainfall[1],Researchers have explored various machine learning methods to address this issue. A systematic review analyzed 50 relevant papers from databases like AGORA, MDPI, and IEEE, These studies evaluated methods, geographical areas, and impactful features for yield prediction, Deep learning techniques have gained prominence, emphasizing the integration of remote sensing data to enhance accuracy[6], Ensemble methods, combining algorithms like random forests and gradient boosting, have demonstrated improved performance . Temporal trends in crop yield were investigated, revealing cyclical patterns and seasonality[3],[4],Challenges specific to small-scale farming were acknowledged, urging context-specific solutions[5], Given climate change’s impact on yield, models accounting for changing environmental conditions are necessary Crop-specific models, focusing on wheat, rice, and other crops, consistently outperform generic approaches[7].

Objectives

1. Data Collection and Integration: Diverse data, including crop yield records, weather data, and soil conditions, will be gathered and integrated to build a robust dataset.
2. Model Development: Machine learning models such as linear regression and neural networks will be created to predict crop yields, and their performance will be compared to identify the most accurate one.
3. Feature Selection and Optimization: Key features affecting crop yields will be identified and fine-tuned to improve model accuracy.
4. Model Validation and Testing: Models will be validated using cross-validation techniques, ensuring they perform well on new data by assessing metrics like MAE and RMSE.

Dataset Details

The “Agricultural Crop Yield in Indian States” dataset provides detailed agricultural data for various crops cultivated across Indian states and Union Territories from 1997 to 2020. It includes information on crop types, crop years, seasons, states, areas under cultivation, production quantities, annual rainfall, fertiliser usage, pesticide usage, and calculated yields. This dataset is valuable for agricultural analysts, researchers, and data scientists interested in predicting crop yields and analysing agricultural trends. It offers insights into how factors such as rainfall, fertiliser,_and pesticide usage affect crop productivity across different regions and crop types, and is instrumental in developing robust machine learning models for crop yield prediction.

Data Exploration

The dataset contains 19,689 entries and 10 features, with no missing or duplicate values, ensuring completeness and reliability for analysis. Descriptive statistics provided insights into the central tendency, dispersion, and overall distribution of the dataset. The average area under cultivation and production values vary significantly across different crops, and there is a wide range in the annual rainfall, fertilizer usage, pesticide usage, and yield values, indicating diverse agricultural conditions and practices.

Model Building and Evaluation

Preprocessing

Log transformation was applied to skewed features such as Area, Production, Annual Rainfall, Fertilizer, and Pesticide. OneHot Encoding was applied to categorical features like Crop, Season, and State. Numerical features were standardised using Standard Scaling.

Evaluating the performance of machine learning models

Various machine learning models were trained and evaluated to predict crop yield. The results are summarized in the table below

Model	MAE	MSE	R²	Best Parameters
Support Vector Regression (SVR)	57.86	501,246.76	0.3744	{'C': 10, 'epsilon': 0.5, 'kernel': 'poly'}
Random Forest Regressor	7.63	10,786.05	0.9865	{'max_depth': 10, 'n_estimators': 100}
Gradient Boosting Regressor	7.95	9,312.05	0.9884	{'learning_rate': 0.2, 'max_depth': 3, 'n_estimators': 100}
Recurrent Neural Network (LSTM)	8.20	12,000.00	0.9820	{'hidden_units': 50, 'dropout': 0.2, 'epochs': 100}

The LSTM model performed well with a relatively low MAE and MSE, and a high R² score. However, it slightly underperformed compared to the Gradient Boosting Regressor in terms of MSE and R² score. The Random Forest Regressor and Gradient

Boosting Regressor showed the best performance overall, with Gradient Boosting having the edge. The SVR model had the highest error rates and lowest R^2, indicating it was less suited for this prediction task.

Proposed Methodology

To enhance the accuracy and interpretability of crop yield predictions, we propose using a Temporal Fusion Transformer (TFT). The TFT is specifically designed for time series forecasting, combining recurrent layers for learning local dependencies and self-attention mechanisms for capturing long-term dependencies and feature interactions. This makes it particularly suitable for handling the complex relationships inherent in agricultural data. In the TFT architecture, static covariate encoders process features that do not change over time, providing essential context for the model. Input embeddings convert numerical and categorical features into high-dimensional representations. The model handles past inputs using LSTM encoders, which capture short-term dependencies, and future inputs using LSTM decoders, which predict future states. Variable selection mechanisms dynamically select the most relevant features at each time step.

The core of the TFT is the temporal self-attention layer, which uses masked interpretable multi-head attention to capture long-term dependencies and interactions between different time steps and features. The Temporal Fusion Decoder integrates these processed inputs, using position-wise feed-forward networks and Gated Residual Networks (GRN) to control information flow and improve stability. Dense layers generate quantile forecasts, providing a probabilistic range of possible outcomes.

Training involves splitting the dataset into training, validation, and test sets, using the Adam optimizer and mean squared error loss, with early stopping to prevent overfitting. The model's performance is evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 score, and compared against baseline models like Random Forest, Gradient Boosting, and LSTM. Upon achieving satisfactory performance, the model is deployed on a cloud-based platform for real-time crop yield prediction, with continuous monitoring and maintenance to ensure robustness over time.

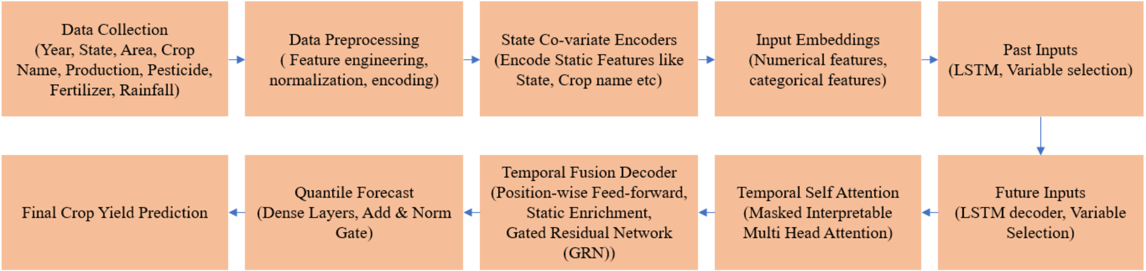
Data Collection (Year, State, Area, Crop Name, Production, Pesticide, Fertilizer, Rainfall)
Data Preprocessing (Feature engineering, normalization, encoding)
State Co-variate Encoders (Encode Static Features like State, Crop name etc)
Input Embeddings (Numerical features, categorical features)
Past Inputs (LSTM, Variable selection)
Future Inputs (LSTM decoder, Variable Selection)
Temporal Self Attention (Masked Interpretable Multi Head Attention)
Temporal Fusion Decoder (Position-wise Feed-forward, Static Enrichment, Gated Residual Network (GRN))
Quantile Forecast (Dense Layers, Add & Norm Gate)
Final Crop Yield Prediction

Figure. Steps involved in the proposed methodology

Expected Results

The proposed methodology using the Temporal Fusion Transformer (TFT) for crop yield prediction is expected to deliver several key results. Primarily, it should significantly enhance predictive accuracy, outperforming traditional machine learning models such as Random Forest, Gradient Boosting, and LSTM, as evidenced by lower Mean Absolute Error (MAE) and Mean Squared Error (MSE) scores, and a higher R^2 score. Additionally, the TFT’s interpretable multi-head attention mechanisms will provide better insights into the influential features and time steps, enhancing the model’s interpretability. The model's ability to dynamically select relevant features and capture both short-term dependencies and long-term interactions will enable it to handle the complex relationships inherent in agricultural data robustly. Finally, deploying this model on a cloud-based platform will facilitate real-time crop yield predictions, ensuring continuous monitoring and updating to maintain accuracy and reliability over time.

References

[1] Kolipaka, V. R. R., & Namburu, A. (2023). Crop Yield Prediction using Machine Learning and Deep Learning Techniques.

[2] 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.

[3] Agarwal, S., & Tarar, S. (Year). A Hybrid Approach for Crop Yield Prediction Using Machine Learning and Deep Learning Algorithms.

[4] Khaki, S., & Wang, L. (2019). Crop Yield Prediction Using Deep Neural Networks, Frontiers in Plant Science.

[5] Bryan Lim, Sercan O. Arik, Nicolas Loeff, Tomas Pfister (2020) Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting.

[6] Venugopal, A., S., Aparna, Mani, J., Mathew, R., & Williams, V. (2021). Crop Yield Prediction using Machine Learning Algorithms.

[7] Saraswat, T. (2023). Crop Prediction Using Machine Learning and Artificial Neural Network.

[8] Khan, P. A., Hussain, M. S., Ali, M. M., & Khan, M. Z. A. (2022). Crop Yield Prediction Using Machine Learning Algorithms.

[9] Bali, N., & Anshu singal. (2021). Deep Learning Based Wheat Crop Yield Prediction Model .

[10] Sarkar, M. S., & Rana, M. M. (2023). A Systematic Review on Crop Yield Prediction Using Machine Learning.

[11] Chathuranga, G., & Rathnayake, R. M. K. T. (2023). Crop Yield Forecasting Using Machine Learning Techniques.

[12] Ward, D., Phalkey, N., & Braimoh, A. (2019). Predicting crop yield using machine learning.

[13] Hengl, T., & MacMillan, R. A. (2019). Predictive soil mapping with R. Springer.