International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Integrated Approach for Crop Yield Prediction in Telangana Region Using Ensemble Techniques and ARIMA Model

P. Sowmya*[1], Dr. A. V. Krishna Prasad[2]

Submitted: 04/02/2024 **Revised**: 12/03/2024 **Accepted**: 18/03/2024

Abstract: This paper presents an integrated methodology for accurate and comprehensive crop yield prediction in the Telangana region, spanning the years 1966 to 2030. Leveraging an ensemble approach, our model combines the strengths of Random Forest Regressor at both the state and district levels, providing granular predictions for each administrative unit. Additionally, we employ an ARIMA model to forecast key meteorological and soil parameters from 2021 to 2030. The ensemble predictions are then integrated with historical data, resulting in a holistic forecast for crop yield. The methodology addresses data sparsity by replacing zeros with mean values, enhancing the reliability of predictions. The proposed approach is validated using robust metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared, demonstrating the robustness and accuracy of the model. The study contributes to the field of precision agriculture, offering insights into the complex dynamics influencing crop yield and providing a valuable tool for sustainable planning in the Telangana region.

Keywords: *ARIMA, Crop yield, Ensemble, Machine learning, Metrics, Random Forest, Telangana.*

1. Introduction

Agricultural planning and resource allocation are critical aspects of ensuring food security and economic stability, particularly in regions with a strong agricultural foundation such as Telangana. Accurate crop yield prediction plays a pivotal role in optimizing these processes. The need for precise forecasts has become increasingly apparent as traditional methods often face challenges in providing reliable estimates. These challenges are compounded by the complex interplay of meteorological, soil, and socio-economic factors that influence crop growth and productivity.

Despite the importance of crop yield prediction, existing methodologies exhibit limitations, such as difficulties in capturing the dynamic nature of agricultural systems and the variability in environmental conditions. Traditional approaches often fall short in providing the granularity required for localized decision-making, especially at the district level. Recognizing these limitations, this research seeks to address these gaps by leveraging advanced techniques in both statistical and machine learning domains.

The motivation behind this research stems from the potential impact on farmers, policymakers, and the agricultural sector at large. Accurate crop yield predictions empower farmers with valuable insights for crop management, resource allocation, and risk mitigation. Policymakers can use this information to formulate targeted interventions, enhance food distribution strategies, and foster sustainable agricultural practices. Consequently, the broader agricultural sector stands to benefit from improved efficiency,

increased yields, and reduced uncertainties.

The main goals of this study include crafting a reliable and precise crop yield prediction model specifically designed for the Telangana region. Leveraging ensemble techniques, specifically the Random Forest Regressor, aims to enhance the precision of predictions at the district level. Furthermore, the research seeks to integrate time-series forecasting methods, such as ARIMA, to model meteorological and soil parameters. This integrated approach aims to address the unique challenges posed by the Telangana agricultural landscape.

This research adopts a multifaceted approach by combining ensemble techniques and time-series forecasting. The Random Forest Regressor is employed to provide district-level predictions, capitalizing on its ability to handle complex, non-linear relationships within the data. Simultaneously, the research utilizes ARIMA models to forecast meteorological and soil parameters over time, ensuring the inclusion of dynamic environmental factors in predictions. This novel integration of machine learning and time-series methodologies enhances the comprehensiveness and accuracy of crop yield forecasts.

This study endeavors to make a meaningful contribution to the agricultural science domain by enhancing the accuracy of crop yield prediction. By addressing the limitations of existing methods and proposing an innovative combination of ensemble techniques and time-series forecasting,

[1] Department of computer Science , Telangana Mahila Viswavidyalayam, koti, Hyderabad, India
[2] Department of computer Science, Maturi Venkata Subba Rao Engineering College,Nadergul, Hyderabad, India
** Corresponding Author Email: Soumya.padam@gmail.com*

anticipates improvements in prediction accuracy and reliability. The application of these methodologies to the Telangana region holds promise for enhancing agricultural productivity, supporting informed decision-making, and ultimately contributing to the sustainable development of the agricultural sector.

To provide a comprehensive exploration of this research, the paper is structured into five main sections. The structure begins with an in-depth literature review, followed by the methodologies, results and discussions, concluding with implications and avenues for future research. This structured approach ensures a thorough examination of the research process, findings, and their implications for the field of crop yield prediction in agricultural science.

2. Related Work

S. Pudumalar et al. introduced a “Crop Recommendation System for Precision Agriculture”, leveraging various machine learning techniques such as Random Forest, CHAID, k-nearest neighbor, naïve Bayes, and ensemble methods. The study incorporates essential soil parameters like depth, texture, pH, color, permeability, drainage, water-holding capacity, and erosion to enhance the accuracy of crop recommendations. Evaluation metrics, namely accuracy score and precision, were utilized to assess the model's performance, resulting in an achieved prediction accuracy of 88%. It was noted that an expanded dataset with a greater number of attributes could potentially contribute to further improvements in accuracy. However, the study's limitations include [limited attribute coverage, regional specificity, Dataset and diversity], emphasizing the need for continued research to address these constraints and enhance the overall robustness of the crop recommendation system.[1]

In their study, Shilpa Mangesh Pande and colleagues introduced a "Crop Recommendation System Employing Machine Learning Techniques." The research leverages Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Artificial Neural Network (ANN), and Random Forest (RF) algorithms, considering diverse factors such as rainfall, temperature, area, season, soil type, year, and season. Evaluation metrics, including mean square error and r2 score, were applied to assess the model's performance. Notably, the Random Forest regressor demonstrated superior accuracy, achieving a remarkable 95%. The study emphasizes the importance of periodically updating datasets for enhanced accuracy and proposes the automation of the prediction process. Limitations are Limited Historical Data, data quality issues.[2]

N. H. Kulkarni and colleagues presented a research initiative titled "Improving Crop Yield Through a Crop Recommendation System Utilizing Ensemble Methods," which utilizes machine learning algorithms such as Naive Bayes and Random Forest.The factors considered encompassed soil quality, rainfall, surface temperature, pH, and sowing season. The ensemble technique showcased remarkable accuracy at 99.91%. Nevertheless, the application of majority voting for ensemble output on extensive datasets revealed challenges, highlighting potential limitations in scalability and computational efficiency.[3]

S. Bangaru Kamatchi et.al. presented a study titled "Enhancing Crop Production through Recommender System and Weather Forecasts," employing artificial neural networks (ANN), a hybrid recommendation system, support vector machines (SVM), and fuzzy c-means. The analysis included parameters such as temperature, precipitation, solar radiation, wind speed, evaporation, and relative humidity. Evaluation metrics included precision, recall, error rate, and accuracy. The hybrid recommendation system exhibited a commendable accuracy rate of 96%. However, potential limitations may arise in terms of computational complexity and system robustness, warranting consideration for further refinement.[4]

Ramesh Medar et.al. introduced a study titled "Enhancing Crop Yield Recommendations through Machine Learning Techniques." The investigation considered a range of parameters such as humidity, rainfall, temperature, cloud cover, soil types (sandy, saline, and peaty), and nutrient levels (copper, potassium, nitrogen, magnesium, iron, calcium) along with soil pH and carbon content. These machine learning techniques aid in determining optimal crop yields. The accuracy of yield predictions can be assessed through various methods, including the implementation of sensor technologies in agriculture. While this research supports informed crop selection for specific lands and seasons, potential limitations may include dependency on accurate sensor data and the need for continuous technology updates.[5]

Dr. Y. Jeevan Nagendra Kumar et.al. presented a study titled "A Supervised Machine Learning Approach for Predicting Crop Yields in Agriculture." The study utilizes machine learning techniques like Random Forest and Decision Tree, taking into account soil health, weather forecasts, past harvests, and environmental factors such as temperature, rainfall, humidity, and soil acidity. The objective is to furnish farmers with actionable intelligence regarding crop demand and pricing trends. However, potential limitations of this approach may include dependency on accurate input data, the need for a robust infrastructure, and the continuous adaptation to changing agricultural conditions.[6]

Sk Al Zaminur Rahman et.al. introduced a study titled "Crop Prediction Based on Agricultural Environment Characteristics using Feature Selection Techniques and Classifiers." The study utilizes machine learning methods like neural networks and Naïve Bayes to forecast crop results using soil and environmental attributes, achieving an impressive accuracy rate of 87.43%. Various evaluation metrics, including accuracy, specificity, recall, logarithmic loss, precision, F1 score, and mean absolute error, were employed to gauge the model's effectiveness. While the results showcase the effectiveness of ensemble techniques in achieving higher prediction accuracy compared to individual classification techniques, potential limitations may include the need for high-quality input data, model interpretability challenges, and the generalization of results across diverse agricultural environments.[7]

Neha Rale et.al. presented a research study titled "Prediction of Crop Cultivation," employing various regression models such as random forest regressor, nearest neighbor regressor, support vector regressor, and gradient boosting trees. The research centers on critical factors like temperature, average wind speed, and precipitation, with root mean square error serving as the metric to assess the model's performance.Despite the valuable insights gained from the models, the study highlights the presence of overfitting, indicating a potential limitation. Despite the valuable insights gained from the models, the study highlights the presence of overfitting, indicating a potential limitation. Addressing this limitation may involve further investigation into model complexity, regularization techniques, or data preprocessing strategies to enhance generalization capabilities.[8]

Anjana et.al. introduced an efficient algorithm titled "An Efficient Algorithm for Predicting Crops Using Historical Data and Pattern Matching Technique." The research employs polynomial regression, random forest, and decision tree models, considering features such as rainfall, temperature, geography of the place, humidity, and soil moisture. The proposed pattern matching method achieves an impressive accuracy of 98%. While the study contributes significantly to crop prediction using various machine learning techniques, a potential limitation could be explored further. For instance, investigating the algorithm's performance under diverse geographical and climatic conditions or addressing the computational complexity of the pattern matching technique may provide valuable insights for future enhancements.[9] [13]

Najla Salah Madlul et.al. present a comprehensive investigation into predicting wheat utilizing the Autoregressive Integrated Moving Averages (ARIMA) model, this research examines crop yields in Iraq from 1988 to 2018. By analyzing annual data, the study identifies the (1, 0, 1) ARIMA model configuration as optimal for forecasting wheat production trends up to 2028.Rigorous statistical tests validate the accuracy of the chosen ARIMA model, projecting a notable increase in wheat production over the next decade. The anticipated annual growth rate surpasses the average observed during the study period (1988-2018), reaching an annual growth average of 0.94%. This research, led by the authors, offers valuable insights that contribute to the field of economic prediction, providing a foundation for well-informed agricultural policies in Iraq.[11]

Dhivya Elavarasan and P. M. Durai Raj Vincent present a groundbreaking contribution to crop yield prediction with their innovative hybrid regression-based algorithm, Reinforcement Random Forest. Centered on tapping into the extensive data repositories publicly accessible across diverse agricultural domains, the algorithm outperforms traditional machine learning methodologies like random forest, decision tree, gradient boosting, artificial neural network, and deep Q-learning. By incorporating reinforcement learning at each stage of attribute selection during tree construction, the algorithm ensures efficient sample utilization. Noteworthy features include a variable significance measure for optimal node splitting, promoting effective training data use. This hybrid approach excels in sparse model structures, requiring minimal parameter tuning, reducing overfitting, accelerating calculations, and enhancing model transparency. Assessment criteria such as Root Mean Squared Error, Mean Squared Error, Coefficient of Determination, and Mean Absolute Error validate the algorithm's exceptional accuracy, achieving an impressive 92.2%.[12]

In their study, K. K. Suresh and S. R. Krishna Priya focus on predicting sugarcane area, production, and productivity in Tamil Nadu using univariate Auto Regressive Integrated Moving Average (ARIMA) models. Covering the period from 1950 to 2007, the dataset undergoes analysis with ARIMA models (1, 1, 1), (2, 1, 2), and (1, 1, 1) identified as suitable for sugarcane area, production, and productivity, respectively. Model performance validation involves comparing predictions with actual values. Subsequently, the developed models are applied to forecast sugarcane area, production, and productivity for future years. Remarkably, the ARIMA models demonstrate effectiveness in short-term forecasting for the agricultural sector at the sub-national level, addressing the asynchronous nature of agricultural cycles across states. This study highlights the critical significance of such short-term forecasting models, particularly in the context of the varying agricultural conditions across regions.[14]

In their investigation, Bushra Praveen and Pritee Sharma explore the repercussions of climate variability on land productivity across major Indian food and non-food grain crops spanning a 50-year period (1967–2016), encompassing 15 crops nationwide. Their study seeks to assess the fluctuations in agricultural output for each crop, considering factors like rainfall and temperature, and also provides forecasts for the next two decades, extending up to 2036. Their analysis reveals a correlation between declining land productivity and increasing mean annual temperatures for most crops, posing a threat to the food security of small-scale and marginalized farming households vulnerable to climate shifts. The study underscores the susceptibility of Indian agriculture to climate change, emphasizing the adverse effects of temperature increases on agricultural yields. Furthermore, utilizing the autoregressive integrated moving average (ARIMA) model for future predictions unveils differing outcomes for various crops. While some crops like gram, sesamum (til), jowar, groundnut, sugarcane, and bajra are anticipated to experience production increases due to expected rises in temperature and rainfall, climate-sensitive crops such as arhar, wheat, rice, cotton, and tea may witness marginal fluctuations in production influenced by temperature changes. [15]

3. Methodology

3.1. Proposed Architecture

The model incorporates historical data obtained from various reputable sources, constituting essential features for analysis. Preprocessing techniques are implemented to improve data quality, followed by training the data using the Random Forest Regressor. The approach adopts a district-wise strategy to account for localized variations effectively. The forecasts generated at the district level are then aggregated to derive a comprehensive prediction for the entire Telangana region. This approach ensures a robust and nuanced understanding of the agricultural landscape, leveraging machine learning techniques to provide accurate predictions for crop yield in Telangana.

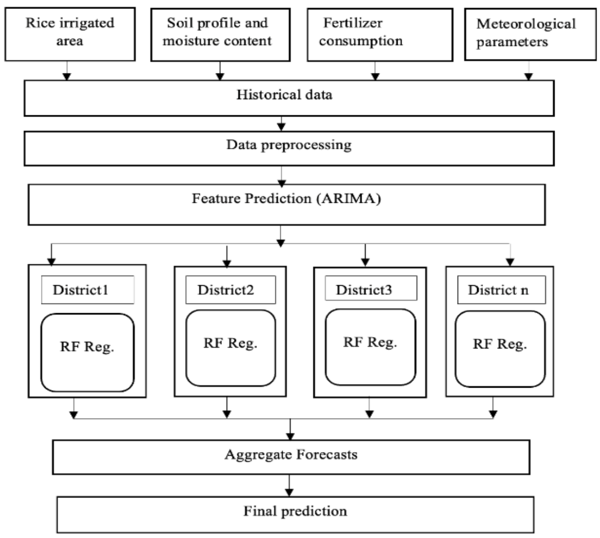

Fig.1. Proposed architecture

3.2. Dataset description

The dataset utilized in this research, denoted as Telangana crop yield dataset, was procured from authoritative sources, specifically from the govt. websites e.g., Telangana open data portal, ICRISAT, nasa.gov etc. The dataset employed in this study, referred to as the Telangana crop yield dataset, was obtained from reputable sources, including government websites such as the Telangana Open Data Portal, ICRISAT, NASA, and others. It encompasses a diverse array of variables crucial to the study, including but not limited to, rainfall, temperature, soil attributes, and crop yield. As part of the data preprocessing phase, rigorous measures were implemented, addressing issues such as missing values through mean imputation. The dataset is structured in the CSV (Comma-Separated Values) format. In this study, the dataset played a pivotal role, serving as the foundation for training machine learning models and facilitating various statistical analyses. It is imperative to note that, owing to the sensitivity of the data, appropriate considerations for data privacy and any imposed restrictions have been duly acknowledged.

Table 1. Sample dataset of Telangana

Dist Code	Year	State Name	Dist Name	RICE IRRIGA	TOTAL CONS	ANNUAL RAI	Precipitation	Max Temp	Min Temp	Wind_Speed	Wind Directi	Surface_soil	Profile_soil_	Root_Zone_	RICE YIELD (Kg per ha)
58	2010	Telangana	Mahabubnag	192.71	143591	755.1	2.45	43.42	9.74	0.06	235.56	0.53	0.64	0.65	2774.63
58	2011	Telangana	Mahabubnag	168.32	136767	494.4	1.26	42.69	11	0.09	254.5	0.43	0.56	0.58	2304.64
58	2012	Telangana	Mahabubnag	137.36	106717	632.7	1.72	44.33	10.56	0.02	266.19	0.41	0.57	0.59	2679.01
58	2013	Telangana	Mahabubnag	164.19	144946	910.8	2.26	44.17	11.73	0.05	246.12	0.52	0.61	0.62	2838.76
58	2014	Telangana	Mahabubnag	157.2	124485	601.1	1.98	42.78	12.7	0.05	206.38	0.48	0.57	0.61	2600.61
58	2015	Telangana	Mahabubnag	92.58	149617	471.9	1.44	44.09	10.93	0.02	215.69	0.4	0.54	0.58	2233.28
58	2016	Telangana	Mahabubnag	139.23	132783	567	1.77	43.75	12.22	0.08	265.38	0.45	0.58	0.6	2696.59
58	2017	Telangana	Mahabubnag	168.46	117612	486	1.89	43.36	12.51	0.06	271.19	0.45	0.58	0.61	2459.12
58	2018	Telangana	Mahabubnag	154.34	116783	831	1.16	44.16	11.31	0.05	266.62	0.38	0.54	0.57	2420
58	2019	Telangana	Mahabubnag	163.67	127658	756	1.57	44.48	10.64	0.04	219.06	0.43	0.55	0.59	2490
58	2020	Telangana	Mahabubnag	145.28	135647	894	3.08	43.56	11.29	0.03	190.12	0.58	0.66	0.69	3380

This dataset, comprising 506 records, offers a wealth of information crucial for unraveling the complexities of crop yield dynamics. Among the key features included are Dist Code, Year, State Code, and Dist. Name, providing the spatial and temporal context for each record. Furthermore, the dataset comprises an extensive range of agricultural and meteorological attributes, such as irrigated rice area (in 1000 hectares), total consumption (in tons), annual rainfall (in millimeters), average precipitation, maximum temperature, minimum temperature, wind speed, wind direction, surface soil wetness, profile soil moisture, and root zone soil wetness. This dataset contains nine district's data.

3.3. District-level forecasting with Random Forest

The dataset is divided into two portions: a training set and a testing set, determined by a cutoff year of 2020. Data up to and including this year are utilized to train the model, while data from the years following 2020 are reserved for evaluating its predictive accuracy. The Random Forest Regressor, initialized with 100 decision trees and a random state of 42, is employed for the machine learning model, ensuring robustness and reproducibility in the forecasting process.

```
# Set the cutoff year for training and testing
cutoff_year = 2020

# Convert the cutoff_year to a datetime object
cutoff_date = pd.to_datetime(str(cutoff_year), format='%Y')

# Split the data into training and testing sets
train_data = telangana_df[telangana_df['Year'] <= cutoff_date]
test_data = telangana_df[telangana_df['Year'] > cutoff_date]

# Initialize the model
model = RandomForestRegressor(n_estimators=100, random_state=42)
```

Fig.2. Random Forest regressor implementation

District-Level Yield Forecasting:

This section outlines the application of Random Forest Regression for district-level rice yield forecasting in the Telangana region. The model is trained for each district, considering key features such as annual rainfall, temperature, and soil characteristics etc. The predictions generated at the district level provide a detailed understanding of rice yield variations across the region.

```
for district in districts:
    # Training data
    X_district_train = train_data[(train_data['State Name'] == state) & (train_data['Dist Name'] == district)][featu
    y_district_train = train_data[(train_data['State Name'] == state) & (train_data['Dist Name'] == district)][targe

    # Testing data
    X_district_test = test_data[(test_data['State Name'] == state) & (test_data['Dist Name'] == district)][features]

    # Convert 'Year' to numerical values for training
    X_district_train['Year'] = X_district_train['Year'].dt.year

    # Train the model
    model.fit(X_district_train, y_district_train)

    # Convert 'Year' to numerical values for testing
    X_district_test['Year'] = X_district_test['Year'].dt.year

    # Make predictions
    y_district_pred = model.predict(X_district_test)
```

Fig. 3. Code for district level yield forecasting

Overall Telangana Forecast:

The subsequent part of the methodology focuses on consolidating district-level predictions to derive an overall forecast for rice yield in the Telangana state. By aggregating the predicted yields from individual districts, this comprehensive forecast offers insights into the broader patterns and trends, aiding agricultural planning and decision-making for the entire Telangana region.

```
# Group by 'Year' and sum the 'Predicted_Yield' for each year
overall_telangana_state_forecast = all_district_predictions.groupby('Year')['Predicted_Yield'].sum().reset_index()
```

Fig. 4. Code for overall Telangana forecast

3.4. District-level District-level agricultural feature forecasting

with ARIMA models

In this phase of the methodology, time series forecasting is employed to predict district-level agricultural features in Telangana. The decision to employ the Autoregressive Integrated Moving Average (ARIMA) model is based on its capability to capture the temporal dependencies inherent in the dataset[10]. The dataset is preprocessed by replacing zero values with the mean, and unique district names are identified. The selected predictors encompass 'CROP IRRIGATED AREA,' 'TOTAL UTILIZATION,' 'ANNUAL PRECIPITATION,' 'Average Precipitation,' 'Maximum Temperature,' 'Minimum Temperature,' 'Wind Velocity,' 'Wind Direction,' 'Surface Soil Moisture,' 'Subsurface Soil Moisture,' and 'Soil Moisture at Root Zone'. The ARIMA model is trained and evaluated for each feature, with predictions extended for the subsequent 10 years. The Root Mean Squared Error (RMSE) is calculated for each district-feature pair to evaluate the precision of the forecasts. Subsequently, the predicted values are utilized for additional analysis and incorporation into the overarching forecasting system.

```
# Iterate over features and fit ARIMA models
for feature_to_forecast in features_to_forecast:
    # Train-test s
    train_size = i (variable) time_series_data: Any
    train, test = time_series_data.iloc[:train_size], time_series_data.iloc[train_size:]

    # Fit ARIMA model
    model = ARIMA(train[feature_to_forecast], order=(3, 1, 2))
    model_fit = model.fit()

    # Make future predictions
    num_future_points = 10
    future_predictions = model_fit.forecast(steps=num_future_points)

    # Append forecasted values to the DataFrame
    forecasted_values_df = pd.concat([forecasted_values_df, pd.DataFrame({
        'Year': range(2021, 2021 + num_future_points),
        'Dist Name': [district_to_forecast] * num_future_points,
        feature_to_forecast: future_predictions.tolist()
    })], axis=1)

    # Evaluate the model
    rmse = sqrt(mean_squared_error(test[feature_to_forecast], model_fit.forecast(steps=len(test))))
    print(f'RMSE for {district_to_forecast} - {feature_to_forecast}: {rmse}')
```

Fig. 5. Code for district-level Agriculture feature forecasting using ARIMA model

3.5. Evaluation for Random Forest regressor for rice yield

prediction in Telangana

In this phase of the study, a Random Forest Regression model is employed to predict rice yield in Telangana based on various agricultural features. The dataset, representing Telangana's agricultural parameters, is preprocessed by replacing zero values with the mean and converting the 'Year' column to numerical values. The Random Forest model is trained using features such as 'RICE IRRIGATED AREA,' 'TOTAL FERTILIZER CONSUMPTION,' 'ANNUAL RAINFALL,' 'Precipitation Avg,' 'Maximum Temperature,' 'Minimum Temperature,' 'Wind Speed,' 'Wind Direction,' 'Surface Soil Wetness,' 'Profile Soil Moisture,' and 'Root Zone Soil Wetness,' while the target variable considered is 'RICE YIELD (Kg per ha)'. The model is evaluated using key performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). The obtained metrics provide insights into the accuracy and predictive capabilities of the Random Forest model for rice yield estimation in the Telangana region.

```
# Calculate evaluation metrics
mae_telangana = mean_absolute_error(y, y_pred)
mse_telangana = mean_squared_error(y, y_pred)
rmse_telangana = np.sqrt(mse_telangana)
r2_telangana = r2_score(y, y_pred)
```

Fig. 6. Code for evaluation metrics for Random Forest regression

4. Results and Discussions

The Random Forest Regression model showed promising performance in predicting rice yield for Telangana, as evidenced by key metrics. Error metrics including Average Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R-squared) were computed to evaluate the model's accuracy. The MAE_telangana of 73.77, MSE_telangana of 10683.66, RMSE_telangana of 103.361, and R-squared_telangana of 0.9791 collectively indicate the model's ability to capture the variability in rice yield. The feature importance analysis identified critical factors influencing predictions, offering insights into the key determinants of agricultural outcomes. Temporal trends in predicted yield over different years were analyzed, revealing patterns and seasonality that can inform strategic planning.

Regional disparities in model performance were also explored, shedding light on variations across districts within Telangana. This nuanced understanding allows for targeted interventions and resource allocation. While the model showcased promising results, it is essential to acknowledge limitations and challenges faced during the modeling process. Factors such as data quality, limited variables, and the dynamic nature of agricultural systems pose challenges that merit consideration.

Table 2. Sample dataset for Mahabubnagar district [2010-2020]:

Dist Code	Year	State Name	Dist Name	RICE IRRIGA	TOTAL CONS	ANNUAL RAI	Precipitation	Max Temp	Min Temp	Wind_Speed	Wind Directi	Surface_soil_	Profile_soil_	Root_Zone_	RICE YIELD (Kg per ha)
58	2010	Telangana	Mahabubnag	192.71	143591	755.1	2.45	43.42	9.74	0.06	235.56	0.53	0.64	0.65	2774.63
58	2011	Telangana	Mahabubnag	168.32	136767	494.4	1.26	42.69	11	0.09	254.5	0.43	0.56	0.58	2304.64
58	2012	Telangana	Mahabubnag	137.36	106717	632.7	1.72	44.33	10.56	0.02	266.19	0.41	0.57	0.59	2679.01
58	2013	Telangana	Mahabubnag	164.19	144946	910.8	2.26	44.17	11.73	0.05	246.12	0.52	0.61	0.62	2838.76
58	2014	Telangana	Mahabubnag	157.2	124485	601.1	1.98	42.78	12.7	0.05	206.38	0.48	0.57	0.61	2600.61
58	2015	Telangana	Mahabubnag	92.58	149617	471.9	1.44	44.09	10.93	0.02	215.69	0.4	0.54	0.58	2233.28
58	2016	Telangana	Mahabubnag	139.23	132783	567	1.77	43.75	12.22	0.08	265.38	0.45	0.58	0.6	2696.59
58	2017	Telangana	Mahabubnag	168.46	117612	486	1.89	43.36	12.51	0.06	271.19	0.45	0.58	0.61	2459.12
58	2018	Telangana	Mahabubnag	154.34	116783	831	1.16	44.16	11.31	0.05	266.62	0.38	0.54	0.57	2420
58	2019	Telangana	Mahabubnag	163.67	127658	756	1.57	44.48	10.64	0.04	219.06	0.43	0.55	0.59	2490
58	2020	Telangana	Mahabubnag	145.28	135647	894	3.08	43.56	11.29	0.03	190.12	0.58	0.66	0.69	3380

Table 3. Features prediction [2021-2030]:

58	2021	Telangana	Mahabubnag	131.42	125567	644.46	2.14	43.69	12.45	0.04	246.09	0.48	0.59	0.62
58	2022	Telangana	Mahabubnag	105.7	136087	632.45	1.82	42.92	11.88	0.05	245.78	0.49	0.58	0.6
58	2023	Telangana	Mahabubnag	94.16	123820	705.11	2.24	42.84	11.91	0.06	244.24	0.48	0.58	0.62
58	2024	Telangana	Mahabubnag	77.76	128673	654.91	2.06	41.93	11.56	0.05	245.18	0.49	0.58	0.6
58	2025	Telangana	Mahabubnag	88.41	129354	654.08	2.25	41.83	11.67	0.05	242.31	0.49	0.57	0.61
58	2026	Telangana	Mahabubnag	90.83	120303	675.73	2.2	40.99	11.37	0.06	242.48	0.48	0.57	0.61
58	2027	Telangana	Mahabubnag	105.48	122575	664.56	2.1	40.84	11.49	0.05	240.6	0.48	0.57	0.6
58	2028	Telangana	Mahabubnag	112.83	121370	662.4	2.17	40.09	11.2	0.05	240.63	0.47	0.57	0.61
58	2029	Telangana	Mahabubnag	118.34	114855	666.73	2.02	39.91	11.33	0.05	239.02	0.47	0.56	0.59
58	2030	Telangana	Mahabubnag	120.84	116049	665.98	2.13	39.25	11.04	0.05	238.84	0.47	0.56	0.6

Table 4. Rice crop yield (District level) [2010-2020]:

Year	State	District	Rice crop yield
2010	Telangana	Mahabubnag	2774.63
2011	Telangana	Mahabubnag	2304.64
2012	Telangana	Mahabubnag	2679.01
2013	Telangana	Mahabubnag	2838.76
2014	Telangana	Mahabubnag	2600.61
2015	Telangana	Mahabubnag	2233.28
2016	Telangana	Mahabubnag	2696.59
2017	Telangana	Mahabubnag	2459.12
2018	Telangana	Mahabubnag	2420
2019	Telangana	Mahabubnag	2490
2020	Telangana	Mahabubnag	3380

Table 5. Rice crop yield prediction (District level) [2021-2030]:

2021	Telangana	Mahabubnag	2594.6271
2022	Telangana	Mahabubnag	2528.0359
2023	Telangana	Mahabubnag	2515.2003
2024	Telangana	Mahabubnag	2489.9973
2025	Telangana	Mahabubnag	2504.2848
2026	Telangana	Mahabubnag	2494.4108
2027	Telangana	Mahabubnag	2537.3646
2028	Telangana	Mahabubnag	2542.3962
2029	Telangana	Mahabubnag	2511.278
2030	Telangana	Mahabubnag	2520.1231

Table 6. Overall Telangana state Rice crop yield prediction [2010-2020]:

Year	Aggregated_yield
2010	28808.04
2011	25245.32
2012	28110.99
2013	28316.88
2014	27664.74
2015	25294.59
2016	31088
2017	27895.17
2018	26964.32
2019	29765.68
2020	32583.54

Table 7. Overall Telangana state Rice crop yield prediction [2021-2030]:

Year	Predicted_Yield
2021	27806.6262
2022	27652.3312
2023	27481.5833
2024	28035.3114
2025	27430.9723
2026	27739.2612
2027	27368.1006
2028	26937.8353
2029	26975.6011
2030	26859.854

5. Conclusion and Future work

In conclusion, the study leverages the Autoregressive Integrated Moving Average (ARIMA) model to forecast all essential agricultural features for all districts in the Telangana region from 2021 to 2030. By employing this time series forecasting approach, the model captures temporal dependencies within the data, providing valuable insights into the expected trends for key agricultural indicators. The methodology involves preprocessing the dataset, replacing zero values with means, and identifying unique district names. The ARIMA model is trained and assessed for each district across all agricultural features, with performance evaluated using the Root Mean Squared Error (RMSE), a measure that quantifies the average difference between predicted and observed values, providing a comprehensive assessment of model accuracy. The predicted values are systematically structured in a

DataFrame, enhancing the insight into projected agricultural trends and supporting informed decision-making in the agricultural sector.

Future research in the field of agricultural forecasting can explore advancements beyond ensemble methods. Scholars could explore the incorporation of cutting-edge technologies like machine learning interpretability tools to improve the transparency and interpretability of predictive models. Furthermore, integrating spatial-temporal modeling techniques may prove advantageous for capturing intricate interactions within agricultural systems, particularly in areas characterized by diverse climate and soil conditions. This approach could contribute to more accurate and localized predictions, catering to the specific needs of different geographic areas.

Another avenue for future exploration involves the integration of remote sensing data and satellite imagery into forecasting models. Harnessing these extensive data reservoirs offers the potential to furnish up-to-the-minute insights into crop vitality, soil moisture levels, and other pivotal determinants, thereby enhancing the accuracy and promptness of forecasts. Investigating the capabilities of hybrid models amalgamating conventional statistical methodologies with cutting-edge machine learning techniques also presents an intriguing avenue for prospective investigations. This multidimensional approach may leverage the strengths of different modeling paradigms, offering a comprehensive solution for agricultural forecasting challenges.

Prospective avenues for research involve delving into supplementary data reservoirs, integrating sophisticated modeling methodologies, and continual fine-tuning to bolster the predictive prowess of the model. The findings of this study carry implications for agricultural strategizing, furnishing stakeholders with actionable intelligence to streamline resource allocation and augment productivity across the region.

References

[1] S.Pudumalar , E.Ramanujam , "Crop Recommendation System for Precision Agriculture", 2016, IEEE Eighth International Conference on Advanced Computing (ICoAC).

[2] Shilpa Mangesh Pande, Prem Kumar Ramesh, Anmol, B R Aishwarya, Karuna Rohilla and Kumar Shaurya "Crop recommendation system using machine learning approach" Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021) IEEE Xplore.

[3] N. H. Kulkarni, G. N. Srinivasan, B. M. Sagar and N. K. Cauvery, "Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique," 2018 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 2018, pp. 114-119, doi: 10.1109/CSITSS.2018.8768790

[4] S.BangaruKamatchi , R.Parvathi, "Improvement of crop production using recommender system,weather forecasts" ,2019,International conference on recent trends in advanced computing.ELSEVIER.

[5] RameshMedar, Vijay S.Rajpurohit, Shwetha, "Crop Yield Recommendation using ML Techniques", 2019, 5th International Conference for Convergence in Technology(I2CT), pp.1-5 ,IEEE

[6] Dr.Y.Jeevan Nagendra kumar, V.Spandana, V.S.Vaishnavi, K.Neha, "Supervised Machine Learning Approach for Crop Yield Prediction in Agriculture sector",2020, Proceedings of Fifth International Conference on Communication and Electronics System (ICCES), pp.736-741, IEEE.

[7] Sk Al Zaminur Rahman, Kaushik Chandra Mitra, S.M MohidulIslam "Soil Classification using machine learning methods and crop suggestion based on soil series",IEEE,2018.

[8] Neha Rale, Raxitkumar Solanki, DoinaBein , James Andro-Vasko "Prediction of crop cultivation" ,IEEE,pp.0227-0232,2019.

[9] Anjana,Aishwarya .K , Ayusha sana ,B Apoorva Bhat,Sharathkumar , Nagaraj Bhat ,"An efficient algorithm for predicting crop using historical data And pattern matching technique ",Received 29 may 2021,Accepted 28 June 2021.(pp.295-297).SCIENCEDIRECT.

[10] Box, G..E., G..M. Jenkins, G..C. Reinsel and G..M. Ljung (2015). The Time Series Analysis:Forecasting and Control, John Wiley & Sons.

[11] Najla Salah Madlul, Eman Younis Al-Najjar, Yusra Tariq Baker and Firas Ibraheem Irhaim ,"*Using the arima models to predict wheat crop production in iraq*" ISSN : 0973-1903, Int. J. Agricult. Stat. Sci. Vol. 16, No. 1, pp. 121-127, 2020.

[12] Dhivya Elavarasan ,P. M. Durai Raj Vincent, "A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters", Journal of Ambient Intelligence and Humanized Computing, Springer-Verlag GmbH Germany, part of Springer Nature 2021.

[13] Rajesh Yamparla; Harisa Sultana Shaik; Naga Sai Pravallika Guntaka; Pallavi Marri; Srilakshmi Nallamothu, "Crop Yield Prediction using Random

Forest Algorithm", 2022 7th International Conference on Communication and Electronics Systems (ICCES).IEEE.

[14] K. K. Suresh, S. R. Krishna Priya, "Forecasting Sugarcane Yield of Tamilnadu Using ARIMA Models", Sugar Tech (Jan-Mar 2011) 13(1):23–26 DOI 10.1007/s12355-011-0071-7.

[15] Bushra Praveen, Pritee Sharma, "Climate variability and its impacts on agriculture production and future prediction using autoregressive integrated moving average method (ARIMA)", Journal of public Affairs", doi:10.1002pa.2016. WILEY.