Optimizing Water Consumption in Crop Production Using Clustering and Machine Learning Models: A Case Study for Arid Regions

[1]Dr. Siddique Ibrahim S P
Assistant Professor
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
siddique.ibrahim@vitap.ac.in

[2]Dhruv Agarwal
UG Student
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
adhruv307@gmail.com

[3]Nikitha Reddy Koya
UG Student
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
nikitha4926@gmail.com

[4]Venkata Sai Srikanth Reddy Venna
UG Student
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
venkatasaisrikanthreddyvenna@gmail.com

[5]Poorna Yoga Saketha Venna
UG Student
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
sakethavenna@gmail.com

[6]Baladithya Devalla
UG Student
School of Computer Science and Engineering
Amaravati, Andhra Pradesh, India
balubaladithya162@gmail.com

***Abstract*— Water is a precious commodity, and as the population grows, its importance increases. To solve this problem, a data-driven approach is used to predict the optimal water use for agriculture. Our focus is Rajasthan, a state of India known for its arid and semi-arid areas, which have severe water scarcity. This research uses machine learning techniques to analyze a variety of data sets, including soil characteristics, crop production, weather and water use. The goal is to determine the water needed for different crops based on soil conditions, climate and available water resources. The results provide practical recommendations for improving Rajasthan's water management and crop selection. The article concludes with future research proposals and agricultural policy and practice implications.**

Keywords*— *crop production; Clustering; Water Management; Machine Learning; optimal water usage; Agriculture; Rajasthan; Irrigation.

I. Introduction

Using an innovative approach that combines machine learning techniques with local environmental data, this study identifies the best water use for a variety of crops in Rajasthan. Unlike conventional methods, this approach uses machine learning techniques to integrate local environmental information. The study combines soil characteristics, crop types and climate data to provide accurate guidelines for water use at the local level, rather than focusing on individual factors such as soil and the weather. By providing a scalable model, the results will encourage sustainable agriculture in other water-scarce regions.

Water scarcity affects the entire world, and agriculture depends heavily on it. Crops require water to grow, and improper management of the resource can reduce crop yield and cause crop death. More than 70% of the freshwater resources in the world are used for agriculture, so maintaining food production depends on water use efficiency.

Once renowned for its arid and semi-arid areas, Rajasthan still struggles with groundwater scarcity, which makes it hard for farmers to decide how much irrigation their fields need. Overwatering frequently results in water loss or underwatering, both of which can harm the crop. In order to make the most of every drop of water, this study looks at sustainable agriculture.

We analyze factors like soil type, minimum water, temperature, moisture and humidity for the best crops grown in Rajasthan. Using synthetic models and machine learning, we predict the optimal water requirement for each product. Based on the availability of water and soil conditions in different regions, we identify the crops that are best for the garden, helping to optimize farmers' yields and make good use of low-water resources.

Our research compiles data from a variety of sources, including local agricultural sectors and seasonal weather data. With machine learning, we aim to improve the accuracy of irrigation methods. We also advocate strategies that can be scaled up to

other water-scarce regions and provide models for sustainable water management. Finally, the goal is to improve crop production while reducing water wastage and addressing economic and environmental challenges in the region.

Figure 1: Rajasthan State Map

II. BACKGROUND STUDY

Although previous research has explored the role of machine learning in agriculture, it still needs to include all aspects of water use. This study aims to address this gap by developing a machine-learning model to predict optimal water use for different crops and regions in Rajasthan.

The purpose of this study is to provide recommendations for proper irrigation based on environmental, agricultural and soil data using synthetic methods and regression models. The findings of this study are important for farmers, agricultural planners and decision-makers to promote better water use and contribute to sustainable agriculture in water-scarce areas.

III. LITERATURE SURVEY

Optimizing water use for crop production is a scientific challenge and an economic and environmental requirement in light of the growing water scarcity. Since agriculture uses a lot of fresh water, sustainable irrigation methods are essential to ensuring that crops receive enough water. These days, cutting-edge techniques like machine learning (ML) and clustering are being used to increase crop yield, improve irrigation efficiency, and support sustainable agricultural practices. This review of the literature examines new developments in using these technologies to maximize agricultural water use.

A study introduced a predictive irrigation scheduling system that employs k-means clustering to establish specific irrigation management zones, taking into account soil hydraulic characteristics and topography. Long short-term memory (LSTM) networks were utilized to forecast soil moisture fluctuations, while a reinforcement learning agent was developed to optimize daily irrigation choices.

Another research introduced a machine learning-based approach combining feature selection methods and stacking ensemble models to determine optimal water quantities for plants. The combined model, which included CART, Gradient Boost Regression, Random Forest, and XGBoost, outperformed the individual models and achieved the highest accuracy and lowest error rate (MSE = 0.0026, MAE = 0.0279, RMSE = 0.0509, R27 = 0.99).

Soil Water Content Prediction: A comparative study evaluated the effectiveness of numerical and ML models for predicting soil water content (SWC). While the HYDRUS-2D numerical model showed the lowest root mean square errors, ML models like ANFIS and SVM performed well, especially under water stress conditions. These models are particularly suitable for SWC predictions when limited data allows for more efficient irrigation planning in resource-constrained environments.

Maize yield prediction: The relationship between irrigation management, spectral groups and maize yield was investigated using ML methods. Random forest models provide high-yield prediction accuracy, especially when combined with spectral and temperature data. This shows the power of integrating irrigation information and management to predict better performance and optimize water use.

Deep reinforcement learning to optimize crop yields: In order to maximize yields while lowering fertilizer and water consumption, a study employed deep reinforcement learning (DRL) algorithms in a crop simulation environment. Finding innovative ways to maximize agricultural yields, satisfy the world's food needs, and preserve water resources has been made possible by this hybrid strategy.

Spatial Optimization of Crop Water Consumption: A model based on cellular automata was used to optimize the spatial distribution of agricultural water usage in the Heihe River basin in China. By significantly boosting regional irrigation benefits by 20.56%, the model improved regional water management and illustrated the need of spatial optimization in water-scarce areas.

Integrating clustering and machine learning models in agricultural practices offers substantial benefits in optimizing water consumption and improving crop yields. These technologies enable precise irrigation scheduling, efficient resource management, and enhanced decision-making, contributing to sustainable agriculture. Future research should focus on refining these models, exploring their scalability across diverse climates and crops, and integrating real-time data from technologies like satellite imaging and Internet of Things (IoT) devices. This will increase the accuracy of water management strategies and ensure their availability in different agricultural environments.

IV. METHODOLOGY

Here, we analyze water use and predict the quantity of water used for different crops. The research will be divided into two main phases – the first thing is data collection, and the second thing is data analysis and interpretation

A. Load the data:

The dataset has 469,833 rows and 19 columns, and this includes info on crop production, water use, soil type and climate in different districts.

B. Pre-process the data:

The quality and consistency of the data across a range of features were ensured by a number of preparatory steps that prepared the data for the development of machine learning models:

1. **Handling Missing Values:** Missing data, especially for temperature and rainfall were filled with the median values.
2. **Encoding Categorical Variables:** As the models need to work better with the data, one hot encoding is used to convert categorical data like crops, and soil types into numerical values.
3. **Feature Scaling:** To avoid any one feature having an excessive impact on the learning process, min-max scaling was used to standardize the values of numerical variables, including temperature, rainfall, and soil moisture, within the range of 0-1.

Min-Max Scaling (Normalization)

When scaling with Min-Max normalization, the data is adjusted to a predetermined range, often [0, 1]. The data scaling formula is

$$X(\text{scaled}) = \frac{(X - X(min))}{(X(max) - X(min))} \quad (1)$$

4. **Outlier Detection:** To avoid distorting the model's predictions, outliers in the water usage data were found and removed.
5. **Data Splitting:** To make it easier to assess model performance and reduce the possibility of overfitting, the dataset was split into training (80%) and validation (20%) subsets.

C. Important Things to Take into Account When Choosing a Technique:

1. **Properties of the Data:** Since the dataset includes a variety of categories (such as weather, soil, and water data), methods like feature scaling and one-hot encoding must be used.
2. **Regional Variability:** As Rajasthan's climate and soil profiles vary greatly among its regions, clustering techniques (such K-means) are crucial for accurately capturing these variances.
3. **Model Accuracy vs. Interpretability:** Accuracy is good for models such as Random Forest and Gradient Boosting, but interpretability is improved by their insights into feature importance.
4. **Computational Efficiency:** Random Forest, K-means, and other models that were chosen are good at handling big datasets.
5. **Scalability:** The methods used ought to be convertible for use in other areas with comparable water scarcity issues.
6. **Data Quality:** The ability of algorithms like Random Forest to handle partial or missing data ensures the robustness of the algorithms.

D. Cluster Forming:

Methods used: Clustering was done using pre-processed items, with encoded categorical variables including climatic data (temperature, precipitation, humidity), soil type, area and statistical data such as water consumption among.

Clustering algorithm: Here, K-Means is used to distribute similar data points based on a number of factors.

The objective is to identify specific clusters in the data set so that they can provide insight into patterns of water use, soil types and crop characteristics.

E. Model training and feature importance:

Models:

A gradient-boosting model was trained to assess the importance of different factors in predicting the water quantity.

A random forest regression model was trained to assess the importance of different factors in predicting water use efficiency.

Analysis:

To evaluate the performance of the model, the mean squared error (MSE) was calculated for the training and test sets.

Identified factors with the most influence on the prediction of the model.

V. RESULTS AND DISCUSSION

We saw different Soil moisture levels in Rajasthan. Bikaner, Jaisalmer and Barmer have higher levels, and Dholpur and Dausa are more arid. This means crops in wet areas require less water, while drier areas like Dholpur require more water.

As a result, by adjusting how the water is used based on particular soil conditions in a particular region, we can manage and conserve water better. And this will help to improve water management and parallelly ensuring the water is used efficiently. By maintaining proper irrigation systems and techniques to specific areas, farmers can reduce water consumption and increase crop yields, ultimately leading to more efficient and sustainable resource allocation to remain permanently in the field.

Highly productive districts: Nagpur (4.28K), Bhilwara (4.00K), and Udaipur (3.98K) show the highest crop productivity, indicating efficient use of water.

Medium manufacturing districts: Ajmer (3.89K) and Sri Ganganagar (3.72K).

The most productive districts are Jaipur (3.15K) and Hanumangarh (3.48K).

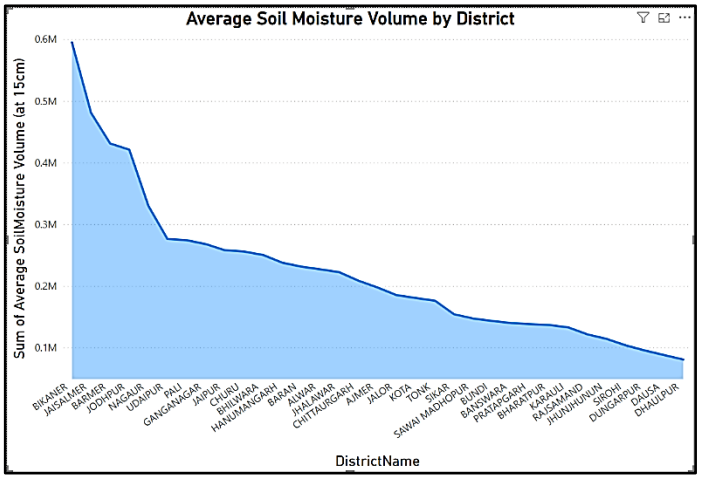

Figure 2: Average soil moisture levels in different districts

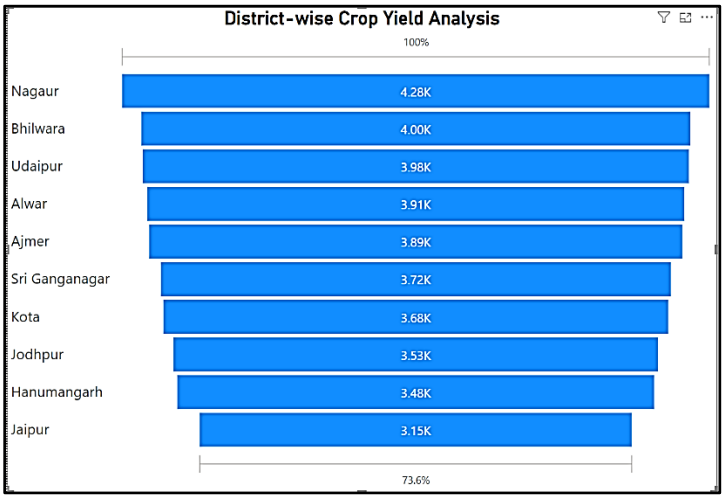

Figure 3: Region-wise crop yield analysis

High water uptake crops: onion (0.63M L/ha) and citrus (0.56M L/ha).

Standard water use: cotton, potato and sugarcane (0.48–0.49M L/ha).

Low-water crops: pomegranate (0.33M L/ha), millet and cumin.

Water availability and utilization: Vegetables like millet, fenugreek and millet use less water. So, they provide opportunities for conservation through better irrigation systems

The higher yields are associated with better water use, while lower yields may indicate water-related challenges. This relationship will be further examined through predictive modelling.

Nitrogen content: Jodhpur has the highest nitrogen content at 3.8K kg/h, and Kota has the lowest at 2.2K kg/ha. These factors will affect the crop water requirements because of the role of nitrogen in plant growth.

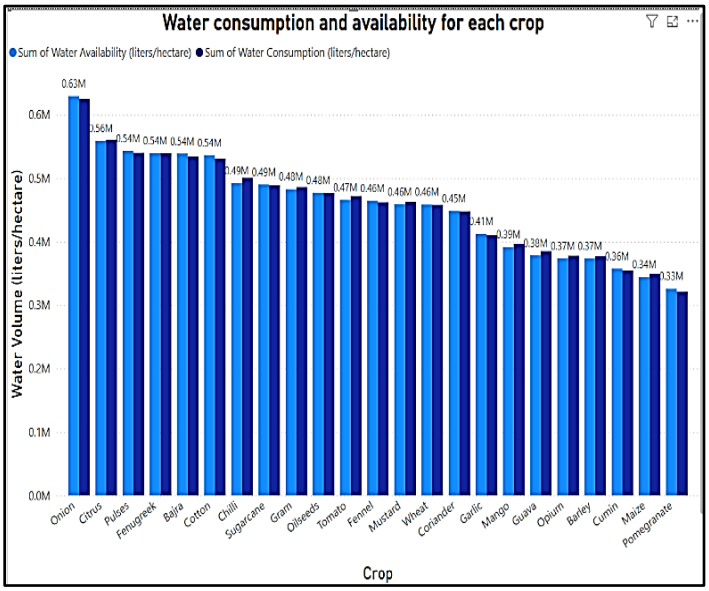

Figure 4: Water consumption and availability of each crop

Potassium levels: Udaipur tops the list with 4.2K kg/ha, and the quota is less than 3.4K kg/ha. Potassium will affect water demand. So, it affects irrigation schedules.

Phosphorus content: Udaipur and Alwar have slightly higher phosphorus content (2.1K kg/ha), and Jaipur has the lowest at 2.0K kg/ha. We know that Phosphorus is important for root growth and nutrient absorption.

Water Management: Nutrient-rich areas like Jodhpur will require a proper amount of water to avoid nutrient losses, and nutrient-poor areas like Kota may require fertilizers and water So that they will irrigate well.

Guidelines: Use equal irrigation in nutrient-rich districts and adjust irrigation in low-nutrient areas to improve crop growth.

Cluster results: size of the cluster:

Cluster 2 shows 187,561 data points

Cluster 0 is showing 155,265 data points and

Cluster 1 is showing 127,007 data points

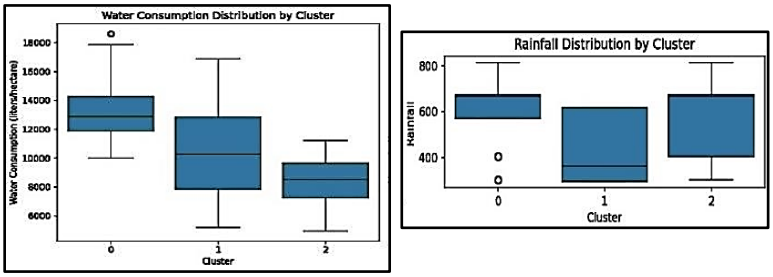

Figure 5: Rainfall and Water consumption by each group

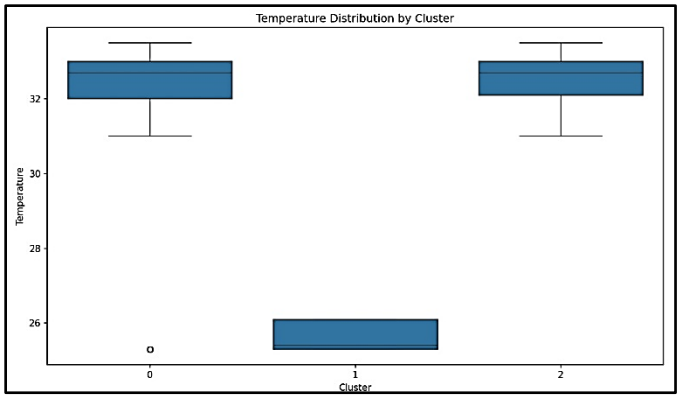

Figure 6: Temperature Distribution by Cluster

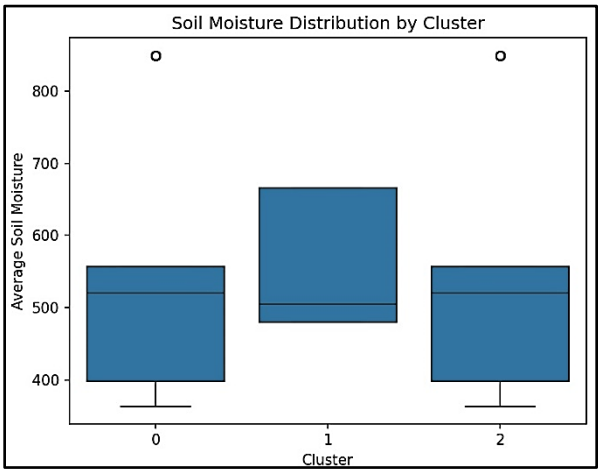

Figure 7: Soil Moisture Distribution by Cluster

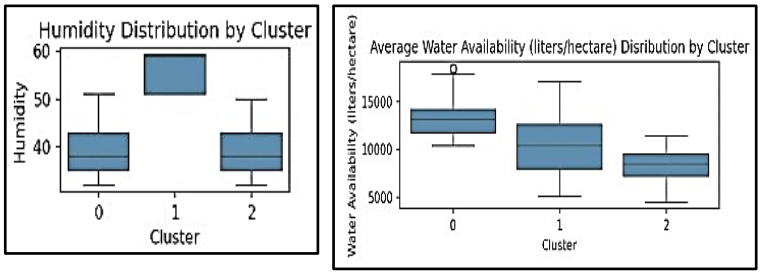

Figure 8: Humidity and Water Availability Distribution in each group

Characteristics of the Cluster:

Cluster 2 represents a group of areas with mild environmental conditions, low humidity, low temperature, and low soil moisture. We team saw that this group has the most data points, meaning that it captures a sample of the data set.

For Cluster 0, it is characterized by high rainfall, high water consumption and low temperature. For cluster 0, it is having better quality of water than cluster 2.

And the Cluster 1 represents the areas with poor climatic conditions and poor soil content.

Importance of the feature:

We saw that availability of water is the crucial feature having importance score of 0.9940.

For the Cluster, importance score is 0.0012.

For the Precipitation thing, Important Score is 0.0005.

For the Crop_Pomegranate, Important Score is 0.0005.

For the rainfall, Important Score is 0.0004.

So here we saw that water availability is the most important factor. agglomeration and moisture also important. This suggests that water availability is an important factor in outcome prediction in the data set, and the clusters provide additional predictive information.

The random forest model got an R2 score of 0.9, which indicates strong predictive accuracy. The water availability significantly influenced forecasts, emphasizing the importance for water management efficiency.

Regional findings: Districts like Jodhpur and Udaipur showed different water consumption patterns because of local climate conditions, indicating the need for local water management.

Crop-based variation: The water use is different from crop to crop. The distinct trends observed for oilseeds and onions. This helps to optimize the efficiency of irrigation systems for specific crops.

Key influencers:

Yield and production are the major drivers of water use, with a mutual information score = 5.03.

Area: Larger areas will generally consume more water.

Rainfall and soil moisture are the Secondary factors. Because the rainfall reduces water demand.

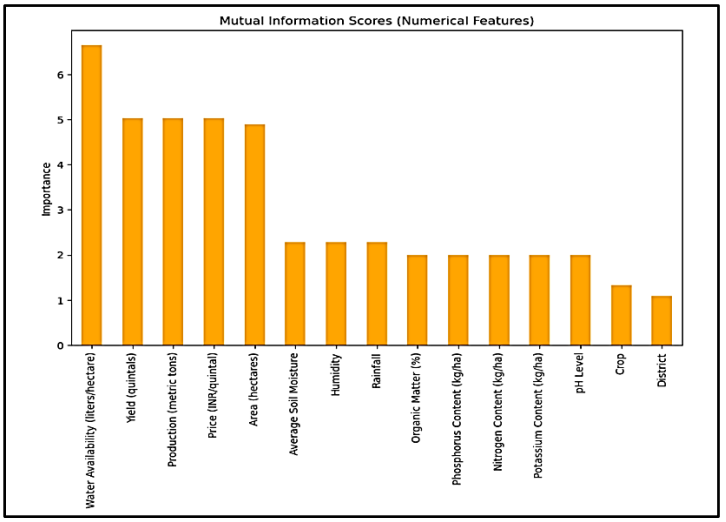

Figure 9: Highlight Significance Based on Common Data Scores for Water Utilization Forecast.

TABLE I. CROP TO BE GROWN BASED ON WATER AVAILABILITY

Optimal Water Usage By District & Crop For Better Yield			
S.No.	***District***	***Crop***	***Predicted Water Consumption (liters/hectare)***
1.	Ajmer	Fennel	10585.552880
2.	Alwar	Fennel	10585.552880
3.	Bhilwara	Fennel	10585.552880
4.	Hanumangarh	Fennel	10585.552880
5.	Jaipur	Fennel	10585.552880
6.	Jodhpur	Fennel	10628.744764
7.	Kota	Sugarcane	10506.904003
8.	Nagaur	Sugarcane	10506.904003
9.	Sri Ganganagar	Sugarcane	10506.904003
10.	Udaipur	Sugarcane	10444.527917

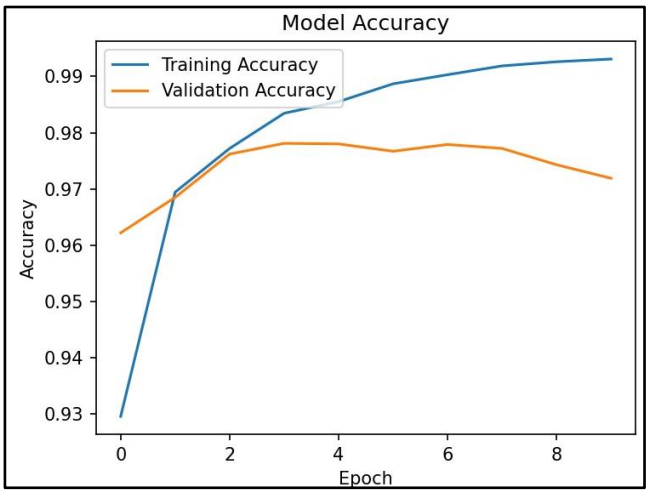

Figure 10: Model Accuracy

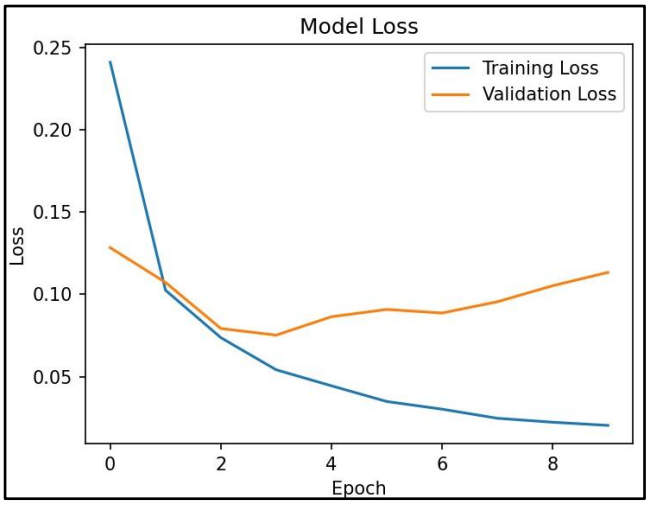

Figure 11: Model Loss

Analyzing Identification Error:

In this study paradigm, identification error refers to errors in classifying crops or locations and predicting the best use of water. Inadequate feature selection, noisy or incomplete datasets, model overfitting, and incorrect data classification can all lead to these errors. As we need to increase the accuracy and make the model to predict more accurate results and reduce errors, it's important to use reliable validation methods, pick the right features, and improve the quality of the data. Reducing errors will also be done by improving the quality of the data, selecting the key features, and by using advanced techniques like hyperparameter tuning and cross-validation. Error reduction may also be facilitated by the incorporation of real-time data and the improvement of clustering techniques.

Increasing the Accuracy of the Model:

Improving the Accuracy of the Model A number of strategies were used in order to achieve a high degree of accuracy in forecasting the ideal water consumption:

1. **Data Preprocessing:** One-hot encoding was used to translate categorical variables, including crop types, missing data were addressed, and feature scaling was used to make sure that every feature had an equal influence on the prediction process.

2. **Feature Engineering:** To understand how the temperature and soil moisture influence water use better, we created new features that combined these two factors.

3. **Model Selection and Tuning:** The study went with random forest and gradient boosting. Because these models are reliable and then used grid search as we need to tune the models for better performance.

4. **Cross-validation:** To reduce overfitting and make sure the model works well on fresh, untested data, a k-fold cross-validation technique (with k=5) was used.

5. **Ensemble Learning:** To improve accuracy by reducing model variance, an ensemble strategy that combined Random Forest, Gradient Boosting, and XGBoost models was applied.

The model's accuracy was significantly improved by these techniques, yielding a cross-validated accuracy of 90% on the valid

We used one hot encoding for categorical data like crop type, soil type etc. to achieve coding process because we need to convert them into a numerical format for model processing.

We can use hyperparameter tuning and cross validation so that the model accuracy will improve. And in parallel it will prevent from overfitting.

We need to Implement real time monitoring, feedback loops to check model predictions and improve them based on actual water usage data to achieve reliability.

VI. RECOMMENDATIONS

Apply Proper Irrigation scheme:- Based on findings Bikaner and Jaisalmer districts have high soil moisture and water availability benefit from precision irrigation methods with different irrigation rates depending on current climate, soil moisture and crop conditions Sensible to locality and their vegetables with appropriate contrast of importance Based irrigation systems should be used to support self-management of

water in reducing water losses and increasing crop production on farms near rivers.

Adopt crop-specific irrigation policies: - It is clear from this study that there are significant differences in water use between crops. Onions and citrus will require different handling methods than less common methods such like corn and cumin. Farmers should use predictions from the model for irrigation based on the type of crop. Where water is plentiful, tobacco can succeed but the other drought-resistant crops grow in soil with less water when it comes to planting.

Use fertilizers and irrigation with appropriate nutrient levels:- The amount of nutrients (nitrogen, phosphorus and potassium) finds the amount of water required. In nutrient rich areas like Jodhpur care should be taken with the water used, to avoid the loss of these nutrients. Precision farming techniques should be used to manage nutrient application and irrigation simultaneously, so that all are integrated in order to support organic crop production and irrigation water efficiency.

Policy options for community water management:- The differences in water consumption levels highlight the local water consumption patterns as seen in different areas of Udaipur falling in similar climatic belts as Jodhpur State and local agricultural authorities exchange their water allocations existing systems to suit the unique conditions of each region with crop characteristics. In districts with excess rainfall including soil moisture, consideration should be given to reducing irrigation water when dryers require more targeted distribution.

Sustainable agricultural practices should include:- Sustainable methods such as irrigation and rainwater harvesting can significantly reduce reliance on external water sources. There is a need to encourage farmers to practice sustainable agriculture in low-rainfall but water-intensive areas. Governments can help farmers adopt irrigation systems and possibly rainwater harvesting methods by providing some subsidies.

Encourage non-use of technology through implementation of government policies:- There are many farmers, especially in rural areas, who do not have enough money to invest in improved irrigation systems. The government should provide financial incentives and support to motivate stakeholders to deploy precision irrigation and machine learning techniques that can help in water management. For example, low-interest loans or direct subsidies can be provided to purchase irrigation equipment or install smart sensors.

VII. CONCLUSION

This study demonstrates the potential of using machine learning and clustering techniques to improve water management in agriculture. This analysis highlights the importance of water availability in determining crop yields and the need for appropriate irrigation practices based on local conditions. Implementing proper irrigation in nutrient-rich areas like Jodhpur and controlling water use in arid areas will greatly improve water conservation. Future work will include extending the model to other regions and incorporating crop price data for broader agricultural recommendations.

REFERENCES

[1] "Optimal water utilization in the state of Odisha using precision agriculture," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Odisha, India, 2023, pp. 1804-1807, doi: 10.1109/ICACITE57410.2023.10182891.

[2] M. R. Barusu, P. N. Pavithra, and P. S. R. Chandrika, "Optimal utilization of water for smart farming using Internet of Things (IoT)," 2023 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2023, pp. 1-7, doi: 10.1109/INOCON53191.2023.9874562.

[3] R. Ben Abdallah, R. Grati, and K. Boukadi, "A machine learning-based approach for smart agriculture via stacking-based ensemble learning and feature selection methods," 2022 18th International Conference on Intelligent Environments (IE), Sfax, Tunisia, 2022, pp. 1-8, doi: 10.1109/IE54923.2022.9826767.

[4] A. Christy Jeba Malar, S. P. Siddique Ibrahim, and M. Deva Priya, "A research cluster-based scheme for node positioning in indoor environment," International Journal of Engineering and Advanced Technology (IJEAT), Vol. 8, Issue-6S, Aug. 2019.

[5] S. P. S. Ibrahim, I. Rakshitha, T. Vasisri, R. H. Aswitha, M. R. Rao, and D. V. Krishna, "Revolutionizing solar generation data mining through advanced machine learning algorithms: Novel insights and results," 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India, 2023, pp. 1-8, doi: 10.1109/CSITSS60515.2023.10334112.

[6] Q. Fang, "Optimizing agricultural water and N managements based on their interactions on crop yield and environment," Agronomy College, Qingdao Agricultural University, Qingdao, China, and Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, China, 2023, pp. 1-10.

[7] S. P. S. Ibrahim, Nithin, S. M. A. Kareem, and G. V. Kailash, "Weed Net: Deep learning informed convolutional neural network based weed detection in soybean crops," 2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, India, 2023, pp. 1-8, doi: 10.1109/ICMNWC60182.2023.10435726.

[8] R. K. Munaganuri and Y. N. Rao, "PAMICRM: Improving precision agriculture through multimodal image analysis for crop water requirement estimation using multidomain remote sensing data samples," School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India, 2023, pp. 1-8.

[9] S. P. S. Ibrahim and M. Sivabalakrishnan, "An evolutionary memetic weighted associative classification algorithm for heart disease prediction," Recent Advances on Memetic Algorithms and its Applications in Image Processing, Part of the Studies in Computational Intelligence book series, Springer (SCI), Vol. 873, pp. 183-199, Jan. 2020.

[10] A. Saad, A. E. H. Benyamina, and A. Gamatié, "Water management in agriculture: A survey on current challenges and technological solutions," Computer Science Department, Oran1 University Ahmed Ben Bella, Oran, Algeria, and LIRMM-CNRS, University of Montpellier, Montpellier, France, 2023, pp. 1-10.

[11] S. Attari, O. Dhatingan, A. Gupta, A. Alshi, and Y. Bais, "Smart AgrIOT: A machine learning and IoT-based complete farming solution," 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022, pp. 1-6, doi: 10.1109/INDICON56171.2022.10039835.

[12] V. Maraš, T. Popović, S. Gajinov, M. Mugoša, V. Popović, A. Savović, K. Pavićević, and V. Mirović, "Optimal irrigation as a tool of precision agriculture," 2019 8th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2019, pp. 1-6, doi: 10.1109/MECO.2019.8760219.