IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 5, May 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305393 |

Empowering Farmers through Crop Yield Prediction using Machine Learning

Shubham Bhandari, Rahul Deokar, Sandip Tarakh, Sanika Hipparagi

Department of Computer Engineering, JSPM's Imperial College of Engineering and Research, Maharashtra, India

ABSTRACT: This survey paper explores the application of machine learning algorithms in predicting crop yields in the context of Indian agriculture, a critical aspect of agricultural management in the region. With India being one of the largest agricultural producers globally and facing challenges such as climate variability and resource constraints, accurate crop yield prediction is of paramount importance. Leveraging machine learning techniques offers promising avenues for enhancing predictive accuracy and efficiency in agricultural yield forecasting specifically tailored to Indian agricultural practices and conditions. This paper provides a comprehensive overview of the various machine learning algorithms employed in crop yield prediction for Indian crops, highlighting their strengths, limitations, and comparative performance. Additionally, it examines the key factors influencing crop yield in the Indian context, data collection methods tailored to Indian agriculture, and challenges associated with implementing machine learning models in this specific domain. By synthesizing existing research findings specific to Indian crop yield prediction, this survey paper aims to offer insights into the current state-of-the-art methodologies and future directions, thereby contributing to the advancement of sustainable agricultural practices in India.

KEYWORDS: Indian crops ,Crop Yield Prediction,Regularization, Machine Learning, Ensemble Learning

I.INTRODUCTION

The prediction of crop yields plays a pivotal role in agricultural decision-making, impacting farmers' productivity, resource management, and economic sustainability. Traditional methods of yield estimation often rely on historical data, expert knowledge, and statistical models, which may lack accuracy and fail to account for dynamic environmental factors. In recent years, the integration of machine learning techniques has revolutionized crop yield prediction by harnessing the power of data-driven algorithms to analyze complex relationships between various agronomic factors and yield outcomes. This paper provides a comprehensive exploration of the application of machine learning algorithms in crop yield prediction, aiming to address the pressing need for reliable forecasting methods in agriculture. By examining the strengths and limitations of different machine learning approaches, as well as the challenges associated with data collection and model deployment, this survey paper seeks to offer valuable insights for researchers, practitioners, and policymakers involved in agricultural innovation and sustainability efforts.

II. LITERATURE REVIEW

Ananthara, M. G. et al. (2013, February) proposed a prediction model for datasets pertaining to agriculture which is called as CRY algorithm for crop yield using beehive clustering techniques. They considered parameters namely crop type, soil type, soil pH value, humidity and crop sensitivity. Their analysis was mainly in paddy, rice and sugarcane yields in India. Their proposed algorithm was then compared with C&R tree algorithm and it outperformed well with an accuracy of 90 percent [2]. Awan, A. M. et al. (2006, April) built a new, smart framework focused on farm yield prediction clustering kernel methodology and they considered parameters like plantation, latitude, temperature and precipitation of rainfall in that latitude. They had experimented weighted k-means kernel method with spatial constraints for the analysis of oil palm fields [3]. Chawla, I. et al. (2019, August) used fuzzy logic for crop yield prediction through statistical time series models.

They considered parameters like rainfall and temperature for prediction. Their prediction was classification with levels 'good yield' , 'very good yield' [4]. Chaudhari, A. N. et al. (2018, August) used three algorithms namely clustering kmeans, Apriori and Bayes algorithm, then they hybridized the algorithm for better efficiency of yield prediction and they considered parameters like Area, Rainfall, Soil type and also their system was able to tell which crop is suitable for cultivation based on the mentioned features [5]. Gandge, Y. (2017, December) used many machine learning algorithms for different crops. They studied and analyzed which algorithm would be suitable for which crop. They have used K-means, Support vector Regression, Neural Networks, C4.5 Decision tree, Bee-Hive Clustering, etc. The factors implying were soil nutrients like N, K, P and soil ph. [6]. Armstrong, L. J. et al. (2016, July) used ANNs for the

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305393 |

prediction of rice yield in the districts of Maharashtra, India. They considered climatic factors namely (considering range) temperature, precipitation and reference crop evapotranspiration. The records were collected from Indian Government repository from 1998 to 2002 [7]. Tripathy, A. K. et al. (2016, July) were same authors who used support vector machines to predict the rice crop yield with same features as the previous paper mentioned [8]. Petkar, O. (2016, July) were also the same authors who applied for SVM and neural networks for rice crop yield prediction proposed a new decision system which is an interface to give the input and get the output [9]. Chakrabarty, A. et al. (2018, December) analyzed crop prediction in the country of Bangladesh where they majorly cultivate three kinds of rice, Jute, Wheat, and Potato. Their research used a deep neural network where the data had around 46 parameters into their consideration. Few of them were soil composition, type of fertilizer, type of soil and its structure, soil consistency, reaction and texture [10].

Jintrawet, A. et al. (2008, May) used SVR model for crops like rice to predict the yield where the model was divided into three steps- predicting the soil nitrogen weight followed by prediction of rice stem weight and rice grain weight respectively. Their factors were solar radiation, temperature and precipitation along with those three steps [11]. Miniappan, N. et al. (2014, August) used artificial neural network in modelling multi-layer perceptron model with 20 hidden layers for prediction wheat yield which considered factors like sunlight, rain, frost and temperature [12]. Manjula, A et al. built a crop selection and to predict the yield which considered various indexes like vegetation, temperature and normalized difference vegetation as factors. They distinguished between climate factors and agronomic factors and other disturbances caused in the prediction for better understanding [13]. Mariappan, A. K. et al. analyzed the data regarding rice crop in the state of Tamil Nadu, India. They have considered factors like soil, temperature, sunshine, rainfall, fertilizer, paddy, and type of pest used and other factors like pollution and season [14]. Verma, A. et al. (2015, December) used classification techniques like Naïve Bayes, K-NN algorithm for crop prediction on soil datasets which constituted nutrients of soil like zinc, copper, manganese, pH, iron, Sulphur, Phosphorous, Potassium, nitrogen, and Organic Carbon [15]. Kalbande, D. R. et al. (2018) used support vector regression, multi polynomial regression and random forest regression for prediction of corn yield and evaluated the models using metrics like errors namelyMAE, RMSE and R-square values [16]. Rahman, R. M. et al. (2015, June) used mainly clustering techniques for crop yield prediction. The paper explained the analysis of major crops in Bangladesh and divided the variables into environmental and biotic variables. The algorithms applied were linear regression, ANN, and KNN approach for classification [17]. Hegde, M. et al. (2015, June) used multiple linear regression and neuro fuzzy systems for predicting crop yield by taking biomass, soil water, radiation and rainfall as input parameters for the research and their majorly concentrated crop was wheat [18]. Sujatha, R., &Isakki, P. (2016, January) used classification techniques like ANN, j48, Naïve Bayes, Random Forest and Support vector Machines. They have also included both climatic parameters and soil parameters as features in their modelling [19]. Ramalatha, M. et al. (2018, October) used a hybrid approach of combining Kmeans clustering and classification based on modified K-NN approach. The data was collected from Tamil Nadu, India where the majorly concentrated crops were rice, maize, Ragi, Sugarcane, and Tapioca [20]. Singh, C. D. et al. (2014, January) developed an application to advise crops which works on selected districts of Madhya Pradesh, India. The user would give input cloud cover, rainfall, temperature, observed yield in the past and the system would predict the yield and Depending on the trigger values set, the crop will be labeled and obtain the results [21].

III.METHODOLOGY

1) Used Libraries:

- Numpy: Numpy is a fundamental library for numerical computing in Python, providing efficient data structures and functions for array manipulation, mathematical operations, and linear algebra. It is widely used for handling large datasets and performing computations with multidimensional arrays, making it indispensable for data preprocessing and analysis tasks.

- Pandas: Pandas is a powerful data manipulation library in Python, built on top of Numpy, that offers versatile data structures and functions for cleaning, transforming, and analyzing tabular data. Pandas' DataFrame object provides a convenient interface for working with structured data, enabling tasks such as data indexing, slicing, grouping, and aggregation. It is commonly used for data preprocessing, exploratory data analysis, and feature engineering in machine learning workflows.

- Matplotlib: Matplotlib is a comprehensive plotting library in Python, widely used for creating static, interactive, and publication-quality visualizations. It offers a wide range of plotting functions and customization options for generating

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305393 |

line plots, scatter plots, histograms, heatmaps, and more. Matplotlib's pyplot interface provides a MATLAB-like interface for creating plots and configuring plot properties, making it suitable for both simple and complex visualization tasks.

- Seaborn: Seaborn is a statistical data visualization library in Python, built on top of Matplotlib, that provides high-level functions for creating informative and visually appealing plots. It offers a streamlined interface for generating complex statistical plots such as scatter plots, box plots, violin plots, and pair plots, with built-in support for categorical and relational data visualization. Seaborn's integration with Pandas data structures and its emphasis on aesthetics and clarity make it a popular choice for exploratory data analysis and presentation.

- Scikit-learn (sklearn): Scikit-learn is a comprehensive machine learning library in Python, offering a wide range of supervised and unsupervised learning algorithms, evaluation metrics, and utility functions. It provides a unified interface for model training, evaluation, and deployment, with support for common tasks such as classification, regression, clustering, dimensionality reduction, and model selection. Scikit-learn's modular design and consistent API make it easy to experiment with different algorithms and techniques, facilitating rapid prototyping and development of machine learning pipelines. Additionally, Scikit-learn includes preprocessing tools for feature scaling, dimensionality reduction, and data transformation, as well as utilities for cross-validation, hyperparameter tuning, and model interpretation. Overall, Scikit-learn is a versatile and user-friendly library that is widely used in academic research, industry applications, and educational settings for building machine learning models and analyzing data.

2) Data Collection:

- The process of data collection involved meticulous sourcing from various channels to ensure a comprehensive and representative dataset for analysis. Government websites provided official statistics and reports, offering insights into agricultural trends and patterns at a broader scale. Kaggle datasets, being a repository of community-contributed datasets, offered additional data sources and perspectives, potentially covering diverse aspects of crop yield prediction.

- Specifically focusing on Maharashtra, data collection extended to different districts and villages within the state. This granularity allowed for capturing localized variations in agricultural practices, soil types, climate conditions, and other relevant factors that influence crop yields. By sourcing data from multiple districts and villages, the analysis could encompass a wide range of agricultural contexts, enriching the predictive models with diverse inputs.

During data collection, the dataset incorporated the following features:

1. State names: To categorize data based on the states within India where the crop yield data was recorded.
2. District names: To provide granularity in geographical representation, identifying specific regions within each state.
3. Crop year: Indicating the time period during which crop yield data was recorded, spanning one agricultural cycle.
4. Season names: Categorizing crop yield data based on prevailing agricultural seasons, such as Kharif and Rabi.
5. Crop names: Denoting the specific crops grown in each agricultural season, reflecting agricultural diversity.
6. Area: Measuring the land area under cultivation for each crop in hectares or acres.
7. Temperature, Wind Speed, Pressure, Humidity: Gathering weather data from meteorological stations to understand climatic conditions during crop growing seasons.
8. Soil Type: Incorporating soil data to understand soil properties and fertility status, influencing crop growth and management practices.
9. Nitrogen (N), Phosphorus (P), Potassium (K) Content:
   - These features represent the levels of essential nutrients present in the soil, namely nitrogen, phosphorus, and potassium, which are vital for plant growth and development.
   - Nitrogen (N) is a key component of amino acids, proteins, and chlorophyll, essential for photosynthesis, protein synthesis, and overall plant growth.
   - Phosphorus (P) is involved in energy transfer, cell division, and the formation of DNA and RNA, playing a critical role in root development, flowering, and fruiting.
   - Potassium (K) regulates water uptake, enzyme activation, and osmotic regulation, contributing to plant vigor, disease resistance, and stress tolerance.
   - Soil nutrient levels influence crop nutrient uptake, nutrient availability, and overall soil fertility, impacting crop yield and quality.
   - The data on nitrogen, phosphorus, and potassium content in the soil provides insights into soil fertility status, nutrient management practices, and the need for fertilization to optimize crop productivity.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305393 |

3) Data Preprocessing:

- Upon initial collection, the dataset likely exhibited various imperfections and inconsistencies that needed to be addressed before further analysis. Common issues included missing values, erroneous data entries, and discrepancies in data types.
- Null values were handled using appropriate techniques such as imputation, where missing values were replaced with estimated values based on statistical measures or domain knowledge. Alternatively, records with missing values might have been removed, depending on the extent and impact of missingness on the analysis.
- Data types were standardized to ensure uniformity and compatibility across different features. This involved converting categorical variables into numerical representations and standardizing numerical variables to a consistent scale. Standardization is crucial for facilitating meaningful comparisons and computations across different variables.

4) Regularization:

- Regularization techniques were employed to address the issue of overfitting, which occurs when a model captures noise or irrelevant patterns from the training data, leading to poor generalization performance on unseen data.
- In the context of crop yield prediction, regularization helps prevent models from becoming overly complex and overly reliant on specific features or data points. By penalizing large coefficients or imposing constraints on model parameters, regularization encourages simpler and more generalized models that are less susceptible to overfitting.
- Common regularization techniques include L1 (Lasso) regularization, which introduces sparsity by driving some coefficients to zero, and L2 (Ridge) regularization, which limits the magnitude of coefficient values. These techniques strike a balance between model complexity and predictive performance, promoting models that are robust and reliable across different datasets.

5) Data Preparation:

- Data preparation involved transforming the cleaned dataset into a suitable format for analysis and modeling. This step typically included feature engineering, where new features were derived from existing ones or domain-specific knowledge to capture relevant patterns and relationships.
- Feature engineering techniques may have included creating interaction terms between variables, aggregating information across multiple variables, or encoding temporal or spatial relationships. These transformations aim to extract meaningful information from the raw data and enhance the predictive power of the models.
- Additionally, data preparation may have involved partitioning the dataset into training, validation, and test sets to facilitate model training, evaluation, and validation. This partitioning ensures that the model's performance is assessed on unseen data, providing a more accurate estimate of its generalization ability.

6) Applying Machine Learning Algorithms:

- Ensemble learning techniques were selected for crop yield prediction to harness the collective intelligence of multiple models and improve predictive performance.

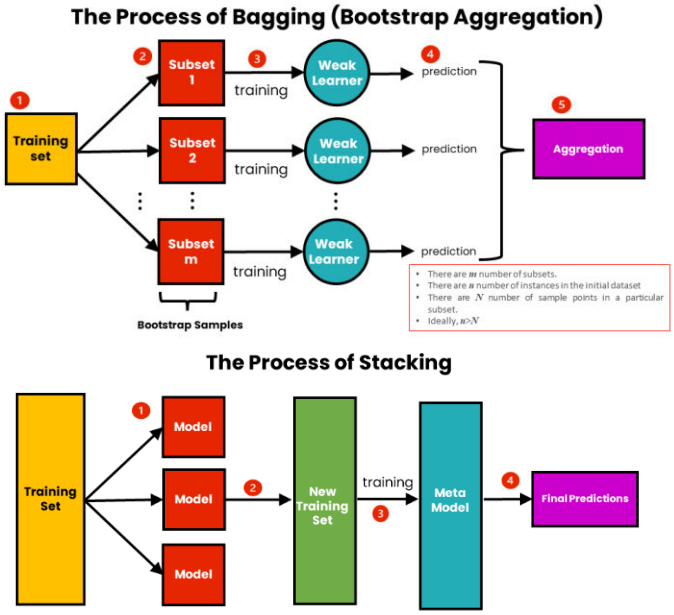
- Ensemble learning methods combine the predictions of diverse base learners, such as decision trees, neural networks, or regression models, to produce a consensus prediction that often outperforms individual models. This diversity in model architectures and learning strategies helps mitigate the risk of overfitting and bias, while also capturing different aspects of the underlying data distribution.

- Bagging (Bootstrap Aggregating), boosting, and stacking are popular ensemble learning approaches, each offering unique advantages in terms of model complexity, interpretability, and robustness. Bagging generates multiple models by bootstrapping the training data and combines their predictions through averaging or voting. Boosting sequentially trains multiple weak learners, focusing on instances that are misclassified by previous models to improve overall performance. Stacking combines the predictions of multiple base learners using a meta-learner, which learns to combine the base learners' outputs optimally.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305393 |



- Ensemble learning is particularly well-suited for crop yield prediction due to the inherent variability and complexity of agricultural systems. By integrating diverse models and learning strategies, ensemble methods offer more robust and reliable predictions, enhancing decision-making in agricultural management.

IV. CONCLUSION

In conclusion, the application of machine learning algorithms holds significant promise for improving crop yield prediction in Indian agriculture. Through the synthesis of existing research and the exploration of various machine learning techniques, it is evident that these methods can offer more accurate and timely predictions, thereby aiding farmers, policymakers, and other stakeholders in making informed decisions. However, several challenges remain, including the need for robust data collection infrastructure, addressing data gaps, and ensuring the scalability and accessibility of machine learning models for smallholder farmers. Future research efforts should focus on addressing these challenges while further refining machine learning models to better accommodate the diverse and dynamic nature of Indian agriculture. Ultimately, by harnessing the power of machine learning for crop yield prediction, India can enhance agricultural productivity, mitigate risks associated with climate variability, and contribute to the sustainable development of its agricultural sector.

REFERENCES

1. “data.gov.in.” [Online]. Available: https://data.gov.in/
2. Ananthara, M. G., Arunkumar, T., & Hemavathy, R. (2013, February). CRY—an improved crop yield prediction model using bee hive clustering approach for agricultural data sets. In *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering* (pp. 473-478). IEEE.
3. Awan, A. M., & Sap, M. N. M. (2006, April). An intelligent system based on kernel methods for crop yield prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 841-846). Springer, Berlin, Heidelberg.
4. Bang, S., Bishnoi, R., Chauhan, A. S., Dixit, A. K., & Chawla, I. (2019, August). Fuzzy Logic based Crop Yield Prediction using Temperature and Rainfall parameters predicted through ARMA, SARIMA, and ARMAX models. In *2019 Twelfth International Conference on Contemporary Computing (IC3)* (pp. 1-6). IEEE.
5. Bhosale, S. V., Thombare, R. A., Dhemey, P. G., & Chaudhari, A. N. (2018, August). Crop Yield Prediction Using Data Analytics and Hybrid Approach. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-5). IEEE.
6. Gandge, Y. (2017, December). A study on various data mining techniques for crop yield prediction. In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)* (pp. 420-423). IEEE.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305393 |

7. Gandhi, N., Petkar, O., & Armstrong, L. J. (2016, July). Rice crop yield prediction using artificial neural networks. In *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)* (pp. 105-110). IEEE.
8. Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016, July). Rice crop yield prediction in India using support vector machines. In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-5). IEEE.
9. Gandhi, N., Armstrong, L. J., & Petkar, O. (2016, July). Proposed decision support system (DSS) for Indian rice crop yield prediction. In 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 13-18). IEEE.
10. Islam, T., Chisty, T. A., & Chakrabarty, A. (2018, December). A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh. In 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 1-6). IEEE.
11. Jaikla, R., Auephanwiriyakul, S., &Jintrawet, A. (2008, May). Rice yield prediction using a support vector regression method. In 2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (Vol. 1, pp. 29-32). IEEE.
12. Kadir, M. K. A., Ayob, M. Z., &Miniappan, N. (2014, August). Wheat yield prediction: Artificial neural network based approach. In 2014 4th
13. International Conference on Engineering Technology and Technopreneuship (ICE2T) (pp. 161-165). IEEE.
14. Manjula, A., & Narsimha, G. (2015, January). XCYPF: A flexible and extensible framework for agricultural Crop Yield Prediction. In 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO) (pp. 1-5). IEEE.
15. Mariappan, A. K., & Das, J. A. B. (2017, April). A paradigm for rice yield prediction in Tamilnadu. In 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 18-21). IEEE.
16. Paul, M., Vishwakarma, S. K., & Verma, A. (2015, December). Analysis of soil behaviour and prediction of crop yield using data mining approach. In 2015 International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 766-771). IEEE.
17. Shah, A., Dubey, A., Hemnani, V., Gala, D., & Kalbande, D. R. (2018). Smart Farming System: Crop Yield Prediction Using Regression Techniques. In Proceedings of International Conference on Wireless Communication (pp. 49-56). Springer, Singapore.
18. Ahamed, A. M. S., Mahmood, N. T., Hossain, N., Kabir, M. T., Das, K., Rahman, F., & Rahman, R. M. (2015, June). Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in Bangladesh. In 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD) (pp. 1-6). IEEE.
19. Shastry, A., Sanjay, H. A., & Hegde, M. (2015, June). A parameter based ANFIS model for crop yield prediction. In 2015 IEEE International Advance Computing Conference (IACC) (pp. 253-257). IEEE.
20. Sujatha, R., &Isakki, P. (2016, January). A study on crop yield forecasting using classification techniques. In 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16) (pp. 1-4). IEEE.
21. Suresh, A., Kumar, P. G., &Ramalatha, M. (2018, October). Prediction of major crop yields of Tamilnadu using K-means and Modified KNN. In 2018 3rd International Conference on Communication and Electronics Syst ems (ICCES) (pp. 88-93). IEEE.
22. Veenadhari, S., Misra, B., & Singh, C. D. (2014, January). Machine learning approach for forecasting crop yield based on climatic parameters. In 2014 International Conference on Computer Communication and Informatics (pp. 1-5). IEEE.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details