JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ADVANCED ANALYTICAL FRAMEWORK FOR CROP YIELD PREDICTION LEVERAGING DIVERSE FEATURE SELECTION METHODS AND MACHINE LEARNING CLASSIFIERS IN VARIED AGRICULTURAL ENVIRONMENTS

[1]Dr. M. Lavanya , [2]K Vinitha, [3]K Vamsi Krishna, [4]K T Siva Kumar, [5]P Maruthi, [6]G Uday Kiran

[1]Associate Professor, [2]20781A3322, [3]20781A3324, [4]20781A3325, [5]20781A3337, [6]20781A3315

[1]CSE(AI&ML)

[1]Sri Venkateshwara College of Engineering & Technology (Autonomous), Chittoor, India

Abstract : In this paper, we propose a prediction method for major crops in India using a combination of K-means clustering and Modified K Nearest Neighbor (KNN) classification algorithms. Given that agriculture is the primary livelihood for over 40 percent of the state's population and the global population is projected to increase by one third between 2010 and 2050, with a corresponding 60 percent rise in demand for crop production, accurate prediction becomes crucial for maximizing yield.

We utilized MATLAB for clustering using the K-means algorithm and WEKA for classification using Modified KNN. The results demonstrate that our method outperforms traditional data mining approaches. This approach allows us to anticipate crop demand more effectively, thereby enabling farmers and stakeholders to optimize production and meet the growing demands of the population."

Index Terms: **Random forest regression, K-means, Machine learning**

I. INTRODUCTION

This paper explores the realm of smart agriculture, aimed at bridging the knowledge gap between traditional and educated farmers by leveraging various data-driven techniques. It focuses on estimating aggregate physical production functions for crop yields in specific states, incorporating technological factors and a newly developed weather index as inputs. Regression analysis, coefficient of determination analysis, and Average Error rate calculations were conducted to compare actual results (target) with predictions from our network outputs.

The primary objective is to develop a user-friendly interface for farmers, providing analysis of rice production based on available data. Different data mining techniques were employed to predict crop yields and maximize productivity. Accurate and timely monitoring of crop conditions and estimating potential yields are crucial for operational programs and decision-making processes in agriculture, especially in countries like Ghana.

The study employs linear regression methods to forecast crop yields across different seasons, recognizing the importance of data mining in various economic sectors. Initially utilized by large companies to analyze consumer data for profitability, data mining methods have expanded into sectors like agriculture and biofuel industries, aiding decision-making processes.

Corn production information is well-established, but various factors like planting date, fertilization, tillage, crop rotation, and weed control practices can influence yield and profitability. The paper also addresses the increasing global energy demand,

focusing on designing and implementing a system to control motor performance using Short Message Service (SMS) via cell phones. This system allows remote control of motor functions and provides status updates via mobile phones, enhancing operational efficiency.

The integration of IoT systems in smart farms enables connectivity among diverse agricultural devices, facilitating intelligent agricultural services based on shared expert knowledge. Additionally, understanding crop price trends is crucial for agri-business profitability, with research focusing on climate-harvest relationships and price forecasting.

Spatial resolution considerations are vital for capturing environmental variability, particularly in mountainous regions, although high-resolution data are limited in availability globally.

II. LITERATURE SURVEY

A. System Architecture

In Figure 1, the raw data consists of the production figures for major crops, rainfall data, groundwater levels, and cultivation areas in India. Considering the high production levels of rice, maize, ragi, sugarcane, and paddy in India, these crops are identified as the major crops of the state. However, since the raw data is incomplete, preprocessing steps are necessary to make it usable.

Data preprocessing involves converting raw data into a meaningful and understandable format. Typically, data preprocessing consists of three main steps: data cleaning, data transformation, and data reduction.

2.1. Data Cleaning: This step involves removing incomplete or erroneous data to ensure the dataset's integrity. Incomplete data can skew analysis and predictions, so it's essential to clean the dataset before further processing.

2.2. Data Transformation: Transformation involves mapping the data into a uniform format. For instance, data may be available at different time granularities such as hourly, monthly, or yearly. Transforming the data into a consistent format, such as year-wise data, enhances its usability and consistency.

2.3. Data Reduction: Data reduction simplifies the dataset by transforming it from an unorganized form into a more manageable and structured format. One common technique for data reduction is clustering, particularly using the K-means algorithm. In K-means clustering, the number of clusters (K) is chosen beforehand. In this case, five clusters are used, representing different levels of production (very low, low, medium, high, very high). After applying the K-means algorithm, the dataset is clustered based on minimum distances, resulting in labeled output. This labeled dataset can then be used for supervised prediction tasks.

Once the dataset is in a supervised format, algorithms such as Fuzzy, KNN, and Modified KNN can be applied for prediction tasks. These algorithms leverage the labeled dataset to make predictions about crop yields, based on factors such as production, rainfall, groundwater levels, and cultivation area.

III. METHODOLOGY

K-Means clustering

The agricultural data is clustered using the K-Means algorithm, which is an unsupervised clustering algorithm. The data is classified into clusters, with 'k' representing the number of clusters. Initially, centroids are assumed to be the first two values in the dataset. Then, the distance between each data point and the cluster center (centroid) is calculated using the Euclidean formula.

$$\text{Distance} = \sqrt{(x2-x1)^2 + (y2-y1)^2}$$

Assign each data point to the cluster center that has the minimum distance from the data point among all the calculated centroids. Then, recalculate the new cluster centers until there is no change in the clusters from the previous iteration. The agricultural input data, such as rainfall, groundwater, cultivated area, and output crop production, are clustered into categories of very low, low, moderate, high, and very high obtained from the output of the K-means algorithm.

IV. IMPLEMENTATION

In this module, a dataset of groundwater levels spanning the past ten years in India is utilized. The data is converted into a dataframe and preprocessed to eliminate records with zero values in all columns. From the columns 'MONSOON', 'POMKH', 'POMRB', and 'PREMON', the average value is calculated. The middle value is determined, and any values below this middle value are considered as zero, while those above are considered as one. These two values are then applied to create a new column called 'class factor'.

These classifications are shown in Table 3 for input data and Table 4 for output crop production.

Figure1: ***WORK FLOW GROUND WATER LEVEL CLASSIFICATION BASED ON KNN MODEL***

Next, KNN classification is employed to predict groundwater levels using the 'class factor' column. For a test run, a K value of 6 is chosen, and the model is predicted. Subsequently, the accuracy of the KNN model is calculated and displayed.

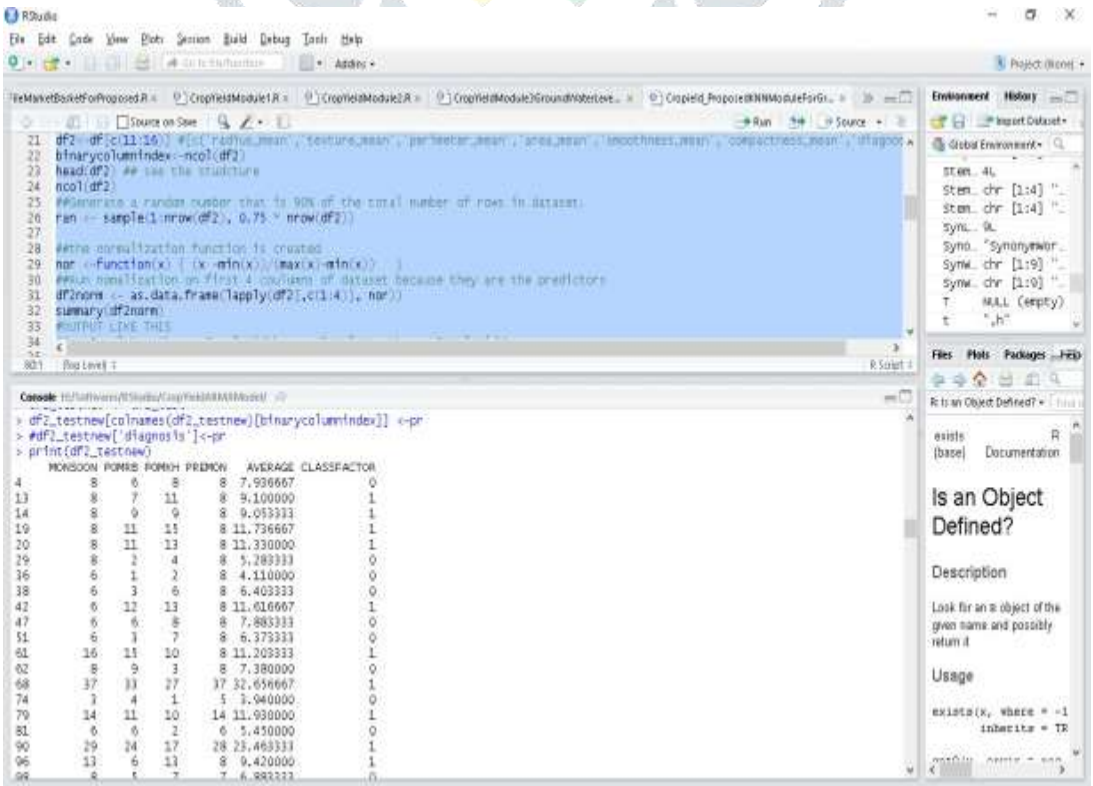
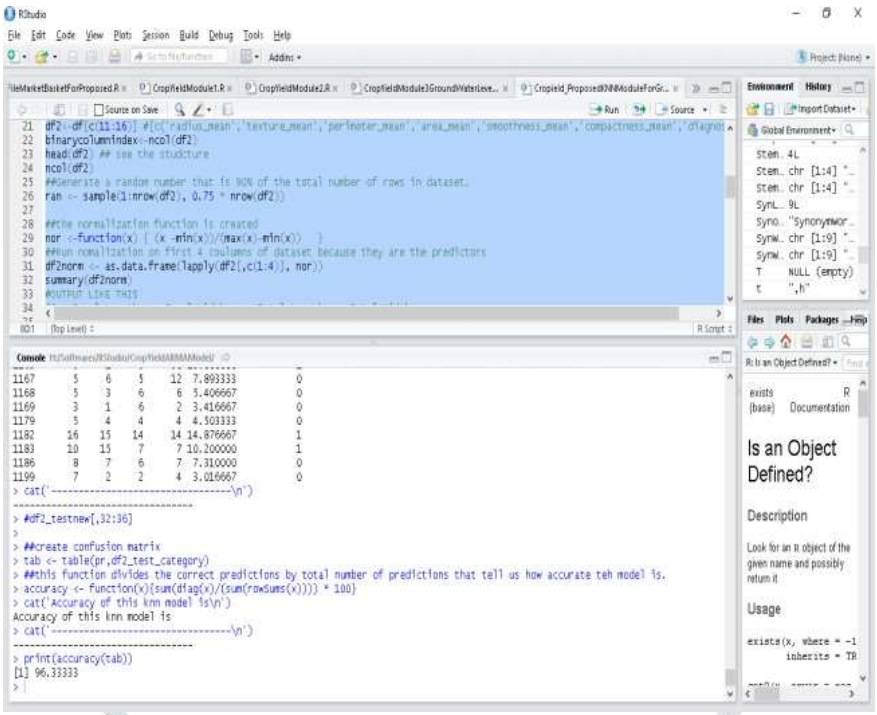


Figure2: KNN CLASSIFICATION FOR GROUND WATERLEVEL DATA



V. ACCURACY OF THE KNN MODEL

5.1 K Nearest Neighbor(KNN)

The nearest neighbor algorithm assigns to a test pattern the class label o fits closest neighbor. Let there be n training patterns, (X1,θ1),(X2,θ2),...,(Xn,θn), where Xi is of dimension d and θi is the class label of the ith pattern. If P is the test pattern, then if d(P, Xk) = min {d(P,Xi)} where i =1...n. Pattern P is assigned to the class θk associated with Xk

Steps involved:

1. Determine the parameter k, which represents the number of nearest neighbors.

2. Calculate the distance between the query instance and all the training samples.

3. Sort the distances and determine the nearest neighbors based on the K-th minimum distance.

In WEKA version 3-6, the training dataset is first selected (Figure 2), and then the corresponding classifier is chosen, with the output being displayed (Figures 3 and 4).

5.2 Modified K Nearest Neighbor

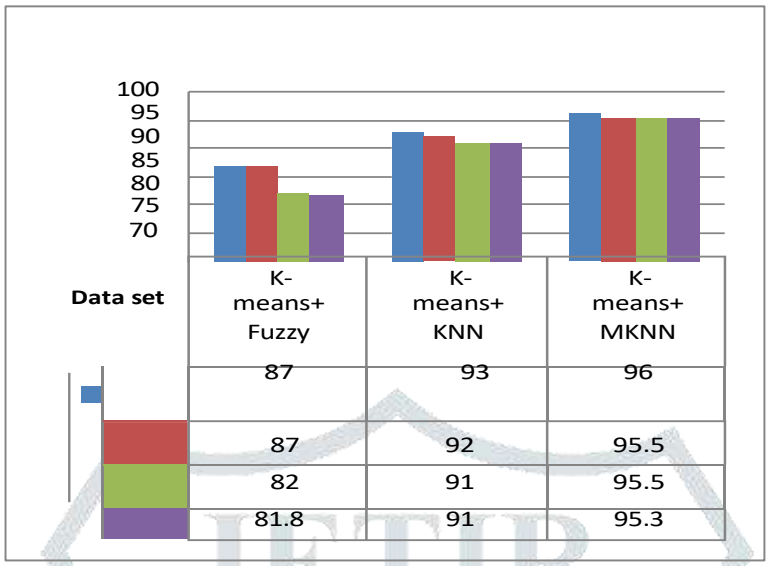
A classification method is proposed to enhance the performance of K-Nearest Neighbor, called Modified K-Nearest Neighbor (MKNN). This method utilizes robust neighbors in the training data. Inspired by this approach, MKNN computes the fraction of neighbors with the same label to the total number of neighbors.

The main idea behind this method is to assign the class label of the data based on K validated data points from the training set. First, the validity of all data samples in the training set is computed. Then, a weighted KNN is performed on any test samples.

VI. RESULT & ANALYSIS

In MATLAB, K-Means clustering was initially performed using parameters such as rainfall, area, and groundwater for analysis. Subsequently, the results from MATLAB were associated with WEKA to conduct further analysis using K-Means with Fuzzy, K-Means with KNN, and K-Means with Modified KNN algorithms. During experimentation with these three algorithms, the precision, measured by correctly classified instances, varied.

After analyzing the output from each algorithm, it was determined that K-Means with Modified KNN yielded the best results among the three experimented algorithms. The analysis results are presented graphically to illustrate the superiority of K-Means with Modified KNN over the other algorithms.



Data set	K-means+ Fuzzy	K-means+ KNN	K-means+ MKNN
	87	93	96
	87	92	95.5
	82	91	95.5
	81.8	91	95.3

The proposed work incorporates fuzzy logic to estimate crop yield, which operates on a set range rather than discrete values. Therefore, errors in predicted rainfall data do not pose significant issues as long as the difference between actual and estimated values is not drastic. The model demonstrates its capability to successfully predict crop yield for a given year when the rainfall and temperature values for previous years are known.

Similarly, the model successfully predicts groundwater levels for a given year when the values from previous years are available. Moreover, the project employs KNN classification to classify groundwater level dataset records, enabling prediction models for future test datasets. This approach facilitates the analysis of past groundwater levels to predict future levels.

In the future, logistic regression can be applied to further classify the data, enhancing the accuracy and reliability of predictions. This integrated approach leverages various techniques to improve the understanding and prediction of agricultural variables, ultimately aiding decision-making processes in agriculture.

VII. CONCLUSION

In conclusion, this study has focused on predicting major crop yields in India using K-Means clustering and Modified KNN classification algorithms. The analysis encompassed three types of algorithms: fuzzy logic, KNN, and Modified KNN. Through rigorous experimentation and evaluation, it was determined that Modified KNN outperformed the other algorithms in terms of accuracy and precision.

Looking ahead, future research endeavors will delve into exploring various bio-inspired methods for crop yield prediction. This entails conducting a comparative study to assess the accuracy and effectiveness of each algorithm in predicting crop yields. By leveraging bio-inspired methods, such as genetic algorithms, swarm intelligence, and neural networks, we aim to enhance the predictive capabilities and robustness of crop yield prediction models.

Ultimately, these advancements hold promise for revolutionizing precision agriculture practices, empowering farmers with more accurate predictions and insights into crop yield variability. By continuously refining and advancing predictive models, we can contribute to the sustainability and efficiency of agricultural production, ensuring food security and prosperity for future generations.

REFERENCE

[1] D Ramesh, B Vishnu Vardhan. "Crop yield prediction using weight-based clustering technique." International Journal of Computer Engineering and Applications, Volume IX, Issue IV, April 15.

[2] Spyridon Mourtzinis, Francisco J. Arriaga, Kipling S. Balkcom, and Brenda V. Ortiz. "Corn yield prediction model uses simple measurements at a specific growth stage." Published on 2 July 2013.

[3] J. Arriaga, Kipling S. Balkcom, and Brenda V. Ortiz. "Corn Grain and Stover Yield Prediction at R1 Growth Stage." Spyridon Mourtzinis, Francisco. Published on May 3, 2013.

[4] Richard Kidd Perrin. "Analysis and prediction of crop yields for agricultural policy purposes." Iowa State University, 1968.

[5] Simone M.S. Costa and Caio A. S. Coelho. "Crop yield predictions using seasonal climate forecasts." Centro de Previsão de Tempo e Estudos Climáticos, Instituto Nacional de Pesquisas Espaciais, Cachoeira Paulista – SP, Brasil.

[6] Askar Choudhury, James Jones. "Crop yield prediction using time series models." Illinois State University.

[7] Mrs. K.R. Sri Preetha M.E., S. Nishanthini, D. Santhiya, K. Vani Shree. "Crop yield prediction." International Journal On Engineering Technology and Sciences – IJETS, March 2016.

[8] Mohammadhossein Hajiyan. "Early Prediction of Crop Yield." School of Engineering, University of Guelph, May 1, 2012.

[9] M. Lavanya, Smart Production and Manufacturing System Using Digital Twin Technology and Machine Learning Springer Nature Journal, 42979-023-01976.