JOURNAL OF BASIC SCIENCE AND ENGINEERING

ENHANCED MODEL TO HANDLE DATA SENSITIVITY FOR CROP YIELD PREDICTION USING MACHINE LEARNING TECHNIQUES

[1]Mrs R.Usha Devi, [2]Dr N A Sheela Selvakumari

[1]Research Scholar, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore.

[1]Assistant Professor, Department of Data Science, Nirmala College for Women, Coimbatore.

[2]Associate Professor, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore

ABSTRACT

The purpose of this study is to gather and analyze data on temperature, rainfall, soil, seed, crop productivity, humidity, and wind speed (in select regions) in order to assist farmers in increasing agricultural yield. Prior to using the proposed architecture, evaluates and processes the massive volume of data, pre-process the data in a Python environment. Second, Dense Region k-Means clustering (DRk-M) is applied to the outcomes of Proposed, yielding an average accuracy result for the data. Additionally, the harvests have been predicted by a self-created recommender system and shown on a graphical user interface created in a Flask environment. The recommended crops of additional states can be found in a similar way in the future thanks to the scalable system design.

Keywords:-Recommendation systems, Clustering, Precision Agriculture, Data Sensitivity

1. INTRODUCTION

The agriculture industry employs the majority of the work force in the nation more than 40% and is the source of direct or indirect dependency for more than 50% of the population [1]. Since agriculture is the foundation of India's economy, it has a significant impact. Abrupt weather patterns cause farmers and agriculture across the nation to suffer since they are unable to produce enough crops [2]. They can no longer support their family and make ends meet, so they are forced to take drastic measures. Further, results in a shortage of food supplies throughout the nation [3]. A more accurate and prioritised predicting system may be created by utilising the large amounts of data generated annually by agriculture, which negates the necessity for the antiquated traditional chart-based prediction systems.

When predicting crop productivity, one can consider the weather to be of utmost importance [4]. The impact of weather on agriculture has been the subject of extensive research; however the majority of these studies call for vast amounts of complex data that are not readily available. This results in the acquisition of data through estimation, which may have beneficial or bad effects. Therefore, to make up for the availability of data, the approach needs to be improved.

1.1.Objective of the study

The primary goal of the study is to forecast weather patterns and assist farmers in making decisions about agriculture in response to those patterns. The study has presented a model to predict the effects of catastrophic weather events and mitigate their effect on global finance, as well as to discover answers to contemporary global issues including global food shortages brought on by frequent climate change. They have created an automated prediction system by utilising Big Data Analytics methodologies. The model is constructed in this study using the Hadoop framework.

1.2.Organization of the paper

The format of this document is as follows. In Section II, current research on big data in agriculture employing a variety of analysis techniques is presented along with recommendations. Section III suggests the system design and methods based on those results. The work is completed in the following section by utilizing pre-existing datasets and implementing the recommender function, clustering algorithm, and Proposed framework to provide the required result. Section IV concludes with discussion of the project's future scope and conclusions.

2. LITERATURE REVIEW

In order to help farmers make decisions about which crops to cultivate and minimise losses from unforeseen or unpredictable calamities, this study discusses how Big Data Analytics, when paired with diverse organised and unstructured data, can be very helpful. Crop yield prediction plays a crucial role in enhancing agricultural outcomes and decision-making processes. Crop yield prediction plays a crucial role in agriculture, aiding farmers, policymakers, and agribusinesses in making informed decisions to enhance productivity and economic outcomes. Various advanced techniques such as machine learning algorithms like Gradient Boosting [6], Stacking Regression [5], and the Crop Yield Prediction Algorithm (CYPA) utilizing IoT techniques [7] have been developed to forecast crop yields accurately. These methods leverage factors like district, area, season, climate, and soil conditions to provide precise predictions, enabling farmers to optimize cultivation practices, resource allocation, and harvest timing. By incorporating big data analysis, decision support algorithms, and active learning strategies, these models achieve high accuracy rates, supporting sustainable agriculture, reducing waste, and improving overall performance in the agricultural sector [8] [9].

Crop yield prediction is a critical aspect of agriculture, especially in India where a significant portion of the economy depends on it [10] [11] [12]. Various factors such as weather conditions, soil composition, and nutrient levels play a crucial role in determining crop output [10] [12]. Machine learning algorithms like random forest, XGBoost, and deep neural networks have been employed to predict crop yields accurately, with reported accuracies ranging from 92.9% to 96% [13]. These predictive models not only help farmers anticipate their harvest but also assist in making informed decisions regarding crop selection, investment planning, and maximizing profitability [14]. By utilizing comprehensive datasets and advanced algorithms, farmers can

optimize their agricultural practices, enhance productivity, and contribute to the overall economic development of the agricultural sector in India [14].

3. MATERIALS AND METHODS

The primary goal, after reviewing the prior research, would be to use Proposed to process the data and create a Python renderer algorithm that would extract output based on the region and seasonal conditions as shown in figure 1. This would be followed by DRk-M to determine the average amount of produce that a group of crops will yield in a given area.

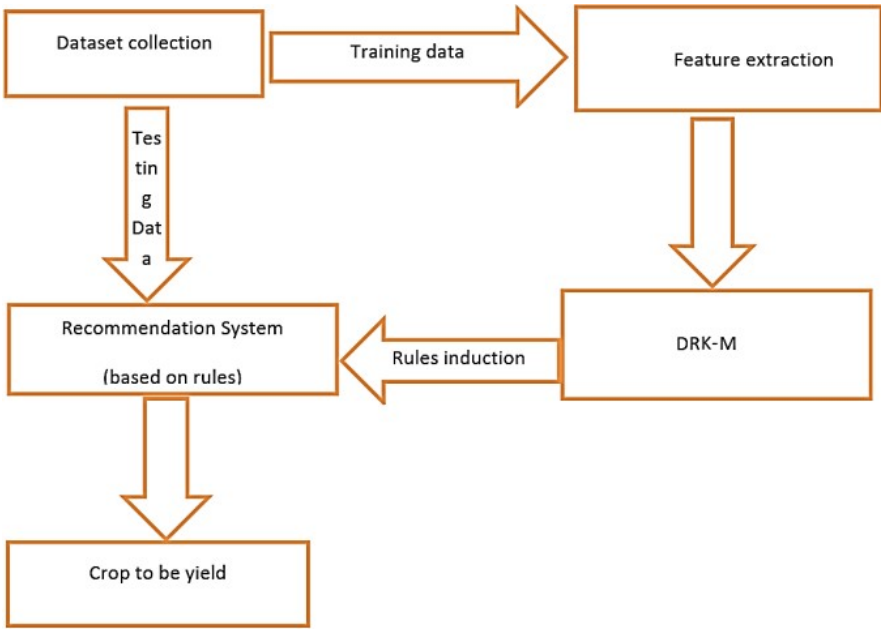


Figure: 1 Proposed Model

3.1.Dataset

The goal of this work is to anticipate crops in India using agricultural and meteorological data. The meteorological data is primarily gathered from publicly available datasets that cover all states' crops, but the agricultural data is limited to three states and two union territories. In this step, several data sets were gathered. Despite a small obstacle, were able to locate seven datasets from Kaggle and agriculture university website that were relevant to workflow and needs. While temperature data dates back to 1995, rainfall data has been gathered since 1901. The crop data was collected starting in 2000 and includes 123 different crops produced in different parts of India. Consequently, the amalgamation of all the data offers a comprehensive perspective of the system and is thus the origin of big data.

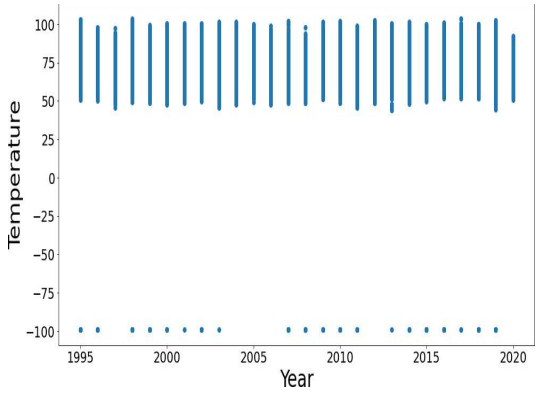
3.2.Proposed DRk-M Methodology

Initially, pre-processed data was imported into HDFS from a variety of sources, including

social media, sensor data, weather forecasts, and so on. HDFS offers backup functionality and dataset storage. The gathered datasets were cleaned and integrated in this instance. Initialized "pandas" data frame to remove the unnecessary columns from our datasets and keep the ones that were significant. A couple of index columns were included for upcoming computations. The estimated value for the missing value in the dataset was statistically determined using interpolation and determined the values of some numerical columns in our datasets that previously had NA values, including the month columns in the rainfall dataset.

dataframe_name.interpolate (inplace = True, method = "linearr, direction = "forwardr") -- (1)

where the forward direction linearly interpolation is carried out and dataframe_name is the name of the dataset being used. The next phase was interpolating the data in a few datasets where it was judged appropriate after the redundant and filthy data was removed using the IQR and z-score method for identifying and removing the outliers. To eliminate the anomalies from the temperature datasets, the Interquartile Range Method outlined was applied. After determining the 25th and 75th percentiles, 1.5 is used as the factor because it is appropriate to take into account only three deviations, or 1.5 multiplied by 2, as anything more or less than these deviations on the sides above and below the 25th and 75th percentiles will result in inaccurate results.



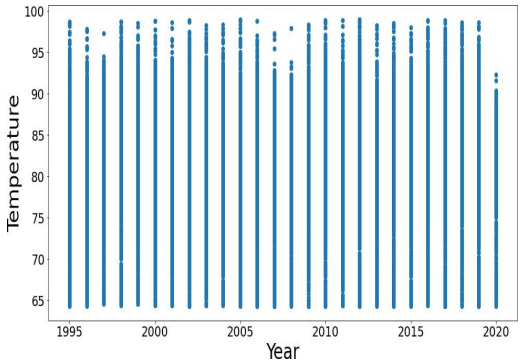
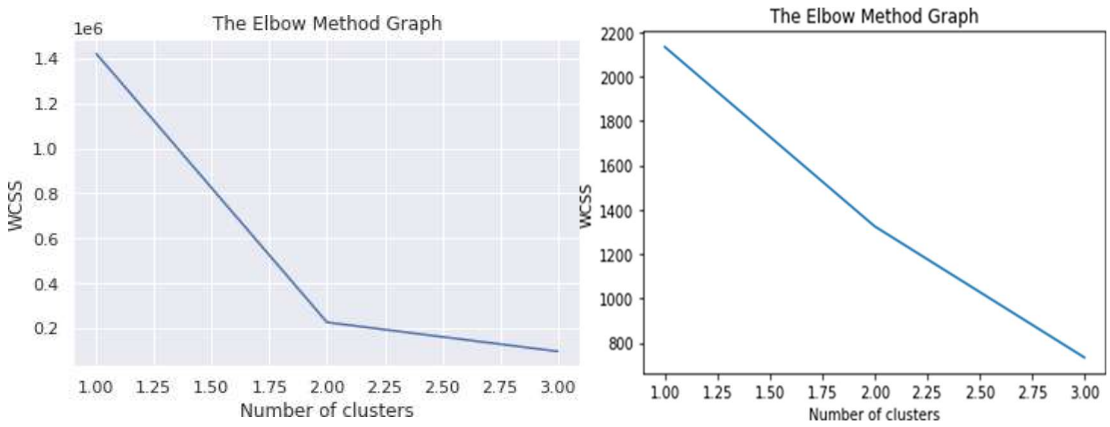


Figure 2. Before InterQuartile range. **Figure 3. After InterQuartile range.**

It takes the desired input from the user and shows them the predicted output, that is, the top three seasonal cropswith the best yield and the top three year-round crops with the best yield along with the expected temperature, rainfall, wind speed and humidity for the input region which gave thatdesired output as show in fig 2 & 3. It also suggests two kinds of suitable soil andthe respective seeds where the crop can be grown.

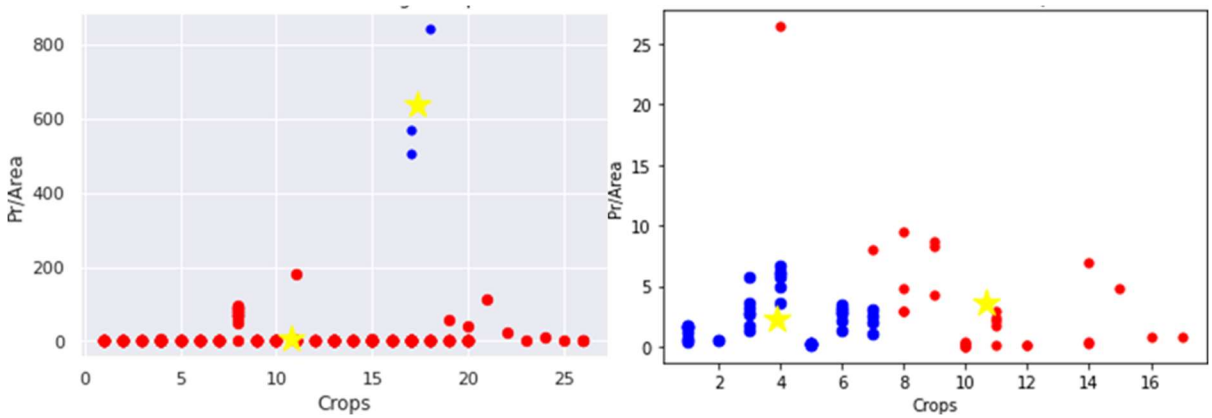
Vol. 20, No. 1, (2023)
ISSN: 1005-0930



(a) (b)

FIGURE 4. (a) &. (b) Elbow graph to find number of clusters

The elbow graph, as shown in Fig. 4a and b, has been created to calculate the number of clusters that, depending on the crops' yield per area, should be built for them in a certain site. The bending point of the graph indicates the optimal number of clusters in the dataset. The elbow of the graph was found in the figures. The crops then plot their produce by area around the cluster centroids to construct clusters using the cluster-forming method.



(a) (b)

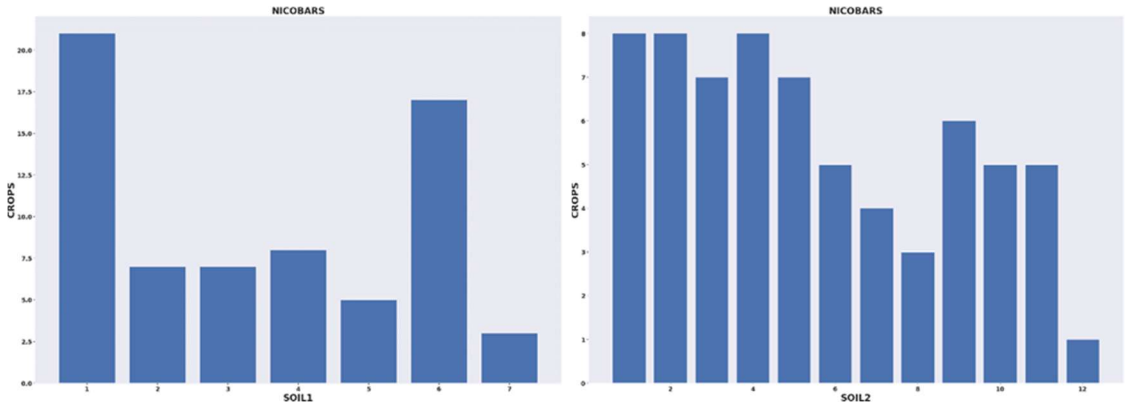
Figure: 5. (a) & (b) Output of DRk-M

To ascertain the correlation between produce per area and a certain crop, a bar graph and scatter plot for a given region are generated using the super dataset. Produces the highest crops per unit area, as seen by the bar graph in Figure 5(a) & (b), the original produce per area and the crops before grouping are plotted to differentiate them before and after the DRk-M process.

Following that, bar graphs were made to investigate the correlation between a soil's

Vol. 20, No. 1, (2023)
ISSN: 1005-0930

appropriateness for a specific kind of seed and that particular seed type. In Figures 6 (a) and (b), the number of seed kinds that can be planted in each of the eight different soil types found in plotted against this number. For every soil type, a number between 1 and 7 has been assigned, and the y-axis shows the maximum number of crops that can be planted in that soil type.



(a) (b)

Figure: 6. (a) & (b) Soil Type and Seeds Grown

The link between different crops, their production per area, and the temperature at which they were cultivated. In some regions, a high produce per area is achieved when the temperature is between 74 and 80 units, however most crops hot region provide a high produce per area if the temperature is between 80.3 and 80.7. The relationship between humidity, crop, and productivity per area. Most crops do well at humidity levels between 74% and 78%, but some demand higher humidity levels between 80% and 84%. The relationship between wind speed and crop and production per area, indicates that certain crops seem to thrive regardless of wind speed fluctuations, whereas most crops need a wind speed of 5-8 units to grow to their full potential. The many elements such as rainfall, temperature, wind speed, and humidity have an impact on the sorts of crops and their yield per area, as demonstrated by the 3D graphs plotted above. It should be emphasised that these factors work in concert to create an environment that is favourable for producing high-quality crop yields per unit area for a significant variety of crops.

4. CONCLUSION AND FUTURE SCOPE

The suggested work use DRk-M to present a crop recommendation system that produces computationally efficient results. A wide variety of crops and their yields per area are the main emphasis of the model, along with the types of soil and seeds that are utilised based on the varieties that are used in a given location. The average yield for a set of crops can be determined using the DRk-M visualisation graphs. As per the methodology, the system is scalable and can be employed to determine the suggested crops of other states in a comparable fashion. If a factor such as wind

speed and humidity is included for every region, this work may be further enhanced to solve the issue of imbalance in the production to requirement ratio and provide a more precise advice. The system's output can be improved by including variables like cloud cover, irrigation, soil moisture, and more. Additionally, the recommender can be altered to provide alerts about potential crop illnesses during a specific growing season and to recommend specific sorts of fertilisers or other nutrients that the soil should contain in order for the crop to thrive and produce its maximum yield.

References:

1. Manida, M., G., Nedumaran. (2020). Agriculture in India: Information About Indian Agriculture & Its Importance. Social Science Research Network,
2. Ohlan, Ramphul. (2018). Agricultural exports and the growth of agriculture in India.. Agricultural Economics-zemedelska Ekonomika, doi: 10.17221/118/2012-AGRICECON
3. Kekane, Maruti, Arjun. (2013). Indian Agriculture- Status, Importance and Role in Indian Economy.
4. P., Sai., Kavita, Bhadu. (2023). Climatic change and its impact of agriculture in India. International Journal of Agricultural Sciences, doi: 10.15740/has/ijas/19.1/387-392
5. N., Chattopadhyay. (2010). Climate Change and Food Security in India. doi: 10.1007/978-90-481-9516-9_15
6. Aravind, Lakshmanarao. (2023). Crop Yield Prediction using Regression Models in Machine Learning. doi: 10.1109/ICAAIC56838.2023.10141462
7. Arirvatham, Mercy, Pushpalatha., P., Kavitha, Rani. (2023). Effective Crop Yield Prediction Using Gradient Boosting To Improve Agricultural Outcomes. doi: 10.1109/ICNWC57852.2023.10127269
8. Fatma, M., Talaat. (2023). Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes. Neural Computing and Applications, doi: 10.1007/s00521-023-08619-5
9. Arya, Phadnis. (2023). Implementation of Prediction of Crop Using SVM Algorithm. International Journal For Science Technology And Engineering, doi: 10.22214/ijraset.2023.52265
10. P., Kalpana., I., Anusha, Prem., S., Josephine, Reena, Mary., ArockiaValan, Rani. (2023). Crop Yield Prediction Using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology, doi: 10.48175/ijarsct-8584
11. Gagan, M., G., N., Reddy., K., S., A., K., K.. (2023). Crop Yield Prediction Using Machine Learning. 1, doi: 10.46632/jdaai/2/1/3

12. Anikó, Nyéki., Miklós, Neményi. (2022). Crop Yield Prediction in Precision Agriculture. Agronomy, doi: 10.3390/agronomy12102460

13. Salvarasan, Iniyan., Vaijinath, A., Varma., Ch, Teja, Naidu. (2023). Crop yield prediction using machine learning techniques. Advances in engineering software, doi: 10.1016/j.advengsoft.2022.103326

14. Anusha, Ashok, Deshmukh., Anushka, Srivatsa., A., A., Arpith, Monteiro., Chaitanya, Gajakosh. (2022). Crop Yield Prediction to Achieve Precision Agriculture using Machine Learning. doi: 10.1109/ICMNWC56175.2022.10031892