IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 20, 2021, accepted September 20, 2021, date of publication October 4, 2021, date of current version October 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3117247

WB-CPI: Weather Based Crop Prediction in India Using Big Data Analytics

RISHI GUPTA[1], (Member, IEEE), AKHILESH KUMAR SHARMA[1], (Senior Member, IEEE), OORJA GARG[1], KRISHNA MODI[1], SHAHREEN KASIM[2], ZIRAWANI BAHARUM[3], HAIRULNIZAM MAHDIN[2], (Member, IEEE), AND SALAMA A. MOSTAFA[2]

[1]Department of Computer Science and Engineering & IT, SCIT, Manipal University Jaipur, Jaipur, Rajasthan 303007, India
[2]Center of Intelligent and Autonomous System, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat 86400, Malaysia
[3]Malaysian Institute of Industrial Technology, Universiti Kuala Lumpur, Bandar Seri Alam, Johor Bahru 81750, Malaysia

Corresponding authors: Akhilesh Kumar Sharma (akhileshshm@gmail.com), Oorja Garg (oorjagarg@gmail.com), and Krishna Modi (krishnamodi2000@gmail.com)

This work was supported by the Universiti Tun Hussein Onn Malaysia under Industry Grant M029.

ABSTRACT This paper aims at collecting and analysing temperature, rainfall, soil, seed, crop production, humidity and wind speed data (in a few regions), which will help the farmers improve the produce of their crops. Firstly, we pre-process the data in a Python environment and then apply the MapReduce framework, which further analyses and processes the large volume of data. Secondly, k-means clustering is employed on results gained from MapReduce and provides a mean result on the data in terms of accuracy. After that, we use bar graphs and scatter plots to study the relationship between the crop, rainfall, temperature, soil and seed type of two regions (Ahmednagar, Maharashtra and, Andaman and Nicobar Islands). Further, a self-designed recommender system has been used to predict the crops and display them on a Graphic User Interface designed in a Flask environment. The system design is scalable and can be used to find the recommended crops of other states in a similar manner in the future.

INDEX TERMS Agriculture, big data analysis, graphical visualization, k-means clustering, map reduce, recommendation system.

I. INTRODUCTION

Due to sudden changes in weather conditions, farmers and agriculture throughout the country suffer as they fail to produce enough crops. This leads them to take serious steps as they are unable to provide for their family and make ends meet. This also leads to a scarcity of availability of food resources in the country. The conditions of farmers in our country need to be changed.

India's economy is greatly influenced by agriculture as it serves as the backbone of the country. More than 50% of the country is dependent directly or indirectly on the agriculture sector and it is responsible for the employment of the major labour force of the country, which accounts for over 40%. Agriculture produces big volumes of data every year, and hence there is a need to get rid of the obsolete traditional predicting methods by charts and use the availability of the big data collected to create a more prioritized and accurate predicting system. Big data will help confront the challenges and enhance the understanding of the whole sector. Big data analytics [34] is the process of examining large data sets containing a variety of data types.

The influence of weather can be deemed as a major priority in the prediction of crop yield. A lot of research work has been conducted in identifying how weather as a factor affects agriculture, but most of these studies require large complex information which is not directly available. This leads to the collection of data by estimation which can have either a negative or a positive effect. Hence improvement is needed in the methodology to compensate for the availability of data.

This work focuses on crop prediction using agricultural and meteorological data in India, which is mainly collected from open dataset sources that contain information of crops from all states, but meteorological data revolves around three states and two union territories. The rainfall data has been collected since 1901, and the temperature data has been found from 1995. The crop data has been amassed from 2000, which comprises the production of 123 crops from various

The associate editor coordinating the review of this manuscript and approving it for publication was Walter Didimo.

regions of India. As a result, the combination of all the data provides an elaborated view of the system and hence serves as the source of the big data. In this project, a MapReduce framework for data processing and a K-means clustering algorithm along with a recommendation function is carried out in the hope to propose crops to sow and elucidate big data applications in agricultural production.

This paper is arranged as follows. Section II presents the existing work done and recommendations in agriculture using big data by various analysis methods. Section III proposes the system architecture and algorithms on the basis of the outcomes defined in section II. In the next section, the work is done using already existing datasets and the MapReduce framework, recommender function and the clustering algorithm is implemented to give the desired output. Finally, in section V, conclusions are made, and the future scope of the project is discussed.

II. LITERATURE REVIEW

A. D. BOSE, "BIG DATA ANALYTICS IN AGRICULTURE" [1]

This paper talks about how Big Data Analytics combined with various structured and unstructured data helps in providing insight to farmers to make a decision as to which crops to grow and reduce losses due to unexpected or unpredictable disasters [23]. In Section I the paper states that we can collect the data produced by sensors from the official databases that are usually maintained and governed by institutions. Here the author suggests we can collect and analyse the data in different stages in agriculture and see their influence in the big picture. It is dependent on two major factors, the push and pull factor [8]. Visualisation of agricultural data is done to simplify the complex, structured, and unstructured data. Interpretation of data can be done using methods like overviews, verifiable models, or in an Ad-Hoc manner and then visualized in the form of tables and graphs [9].

In Section II, the paper talks about techniques like Predictive analysis where we can make the appropriate prediction of the future outcome on the basis of the previous data [33]. A recommendation system is an informative system whose task is to offer an output that is based on functional patterns and behavioral data. Recommender systems generally give useful advice as the output is based on the approach used and the categories. The next method is Data Mining which can be defined as the process of extracting the previously unknown and useful information from large quantities of incomplete data for practical application [35]. It plays a vital role in the agriculture sector, especially discovering patterns in big datasets, i.e., pattern mining Next, the spike and slab regression analytic technique is discussed where the term spike and the term slab are used as a type of coefficient for regression [24]. In the time series analytic technique using big data, time is taken as a variable that is independent with a motive to vegetation price movement, forecast crops and price fluctuation in the current market.

In Section III, the implementation of analytic techniques in agriculture had been discussed. The first method is an Intelligent crop recommendation system that considers all the factors such as soil conditions, temperature, rainfall and location. This system is further split into two different systems: the crop predictor, whose main task is to help agriculturists by recommending crops and the rainfall prediction system that predicts the occurrence of rainfall for each month across the year [17].

The next method discussed was Precision Agriculture using Map-Reduce used to allow variable rates and inputs which help in the understanding of time and space variability in criterion [18]. Here the data is obtained and pre-processed. Then map-reduce is performed, and 3D visualization is done to visualize the output.

Further crop prediction using various machine learning approaches were discussed. A few of them were 1) Grey wolf optimisation (GWO) technique 2) K-means clustering 3) Apriori algorithm 4) Naive Baye. Next Smart Farming was discussed where a few of the services like Internet of Things, Cloud Computing, Mobile Computing were detailed about.

The Crop analysis using Data mining techniques discussed is aimed at analysing greenhouse crops with the help of data mining techniques to extract patterns. With the help of the user interface and selection of specific greenhouse attributes, farmers will be able to predict yield patterns, crop patterns and further make important decisions based on them.

Lastly, the author talks about a Spark-based system to perform collection, learning, training, validation and visualization of distributed data. This method of data analytics can be used for crop yield prediction, current weather trends and performing insights on Agricultural market data [10]. In Section IV, the challenges that are faced in the analysis of big data in agriculture are discussed. The author states that obstacles faced for agriculture are usually Technical or Organizational problems. The paper further mentions the problems faced in the big data analysis of agriculture data, majorly, availability, accessibility and scalability of data for analysis.

Section V talks about the future scope of work where the author goes ahead and discusses various factors that could be helped with like product traceability, genetic engineering, supply chain, yield production, high precision, scientific simulations and so on and so forth. Lastly, Section VI contains a comparison table of big data techniques where one can notice that it suggests that we use MapReduce for weather and climate data and K-Means Clustering for crop and vegetation data by collecting historical datasets.

B. R. PRIYA, D. RAMESH, E. KHOSLA, "CROP PREDICTION ON THE REGION BELTS OF INDIA: A NAÏVE BAYES MapReduce PRECISION AGRICULTURAL MODEL" [2]

In Section I of the paper, the focus is on the system of agriculture in Telangana. The data is collected from Cridas and farms of Hyderabad and Hayathnagar. A recommendation system recommends which crop to cultivate in the related seasons

using Naïve Bayes classifier. Rice, Cotton, Maise and Chillies are the crops taken into consideration.

Section II talks about the previous work done in the field of precision agriculture. The author tells the advantages and the grey areas of methods and models used in previous work like linear regression with neural networks, MapReduce, KNN algorithm, a crop growth prediction model, sequential data assimilation. In Section III, the author describes the proposed methodology which is used to predict which of the four crops are suitable in Telangana. He talks about the modality and methodological conditions of 3 zones, i.e. (i) Northern Telangana, (ii) Central Telangana and (iii) in Southern Telangana, with seven major types of soil in which farmers mainly cultivate soybean, maise, rice, cotton where the water for irrigating the soil is provided by the rivers the Godavari and Krishna and monsoons (June-September). The suitable conditions for growing rice, maise and chillies are discussed. After collecting data from various sources like sensors from fields, images from satellites, data of crop, irrigation reports and weather data, it was pre-processed to find out the missing values and impute them using the mean method. Then feature selection and data extraction were performed in terms of soil, temperature, rainfall and atmospheric pressure. Further MapReduce was implemented on this data, and then a Naive Bayes classifier model for crop prediction was made using the Naive Bayes algorithm [11]. This model recommended two or more crops based on the input data supplied.

Section IV describes the results and recommends sowing and harvesting suitable crops.

1) It was concluded that cotton should be planted in March/April as July to September are its ideal growth months where maximum growth is noted in the month of August. Since there is no noticeable growth from October to December, the crop can be harvested in January or February.

2) Rice is grown in Rabi and Kharif season. It should be sown in July, as there is notable growth from August to September and it can be harvested in October and November.

3) Chilli requires good rainfall; hence the crop is sown at the start of July.

4) The Maize plant should be sown at the end of June as it has the highest growth in July and is to be harvested in preferably September.

Section V discusses possible future enhancements. This work used Naïve Bayes to introduce a crop recommender system to make it very efficient when it comes to computation. The system can be used on a variety of crops as it is scalable.

C. WU FAN, CHEN CHONG, GUO XIAOLING, YU HUA, WANG JUYUN, "PREDICTION OF CROP YIELD USING BIG DATA" [3]

This paper discusses crop yield prediction, food security, Map Reduce and nearest neighbour modelling in terms of big data using agricultural data in China. In Section I, the paper talks about food security and its aspects like producing enough food and maintaining a stable supply of food in the market and how big data [25] can help sort this out and points out that the earliest time in advance and accuracy are the priorities of predicting the crop yield.

Section II portrays the advancement and application of big data in crop yield prediction. The paper states that effective plans for improving the performance of prediction of crop yield and the methods to take the maximum advantage of huge datasets related to agriculture and food security. Currently, big data can be obtained in semi-structural or non-structural forms from Recognition technology, Radio Frequency Identification, Remote Sensing, Weather stations. It is further reviewed by the paper that crop yield forecast is the most addressed topic, followed by climate change impact assessment and water resources. The well-developed methods have been categorised by Statistics methods, Remote Methods, Crop growth simulation, Econometrics. Section III proposes a model based on prior structure and weather data processing structure. First, the data was prepared by collecting it from the China Meteorological Administration with high accuracy of above 99%. Then MapReduce was performed by partitioning the data into multiple sections. Then the map was executed according to certain rules followed by the Reduce function, where data having the same year was rearranged and combined. Output was written to distributed file systems. After that, weather similarity (defined by weather distances) [26] was checked using nearest neighbor's. The smaller was the distance between the two years which was quantified the similar the two years would be. At last, the autoregressive moving average model was used by combining two models. The output produced by one that has white noise as it is input which means that it has a linear relationship. In Section IV, the experiment is conducted on the already existing weather datasets, and the advantages of using this new method are talked through. This crop yield predictor is an application that has its basis on a processing structure that manages data in sequence to search for similar years.

In the first step, the weather data was processed using MapReduce, keeping precipitation, the intensity of sunshine and temperature at ground level as variables to calculate the daily mean and monthly mean. The process was divided into three steps, i.e. Map(to calculate monthly mean value), Reduce (to combine intermediate data), and storing the result. Next, a search for similar years was performed by conducting normalisation on three matrices to obtain a single 59×36 matrix, and then the difference of distance was obtained by computing the norm of the target year. After sorting, 20 nearest neighbours were obtained similar to the target year. This was followed by preparing an ARMA model for prediction based on nearest neighbours found. This model was used to predict the crop yield of 2013 as an example and had a deviation of only 0.5%. The nearest neighbour's method using MapReduce weather data processing structure had a balance of both accuracy and time in advance.

Finally, in Section V, a conclusion is made that using the method mentioned above, an advantage of the already existing large datasets can be taken and put into use. Future possible work includes the faster accumulation of data and

integrating weather calculation into the section that is processing data to reduce computing time. Lastly, this paper importantly focused on data mining in agricultural data from the perspective of time using a time aspect, the MapReduce weather data processing structure. The same methods can be applied to different geographical aspects.

D. M. G. RAMYA, C. BALAJI, L. GIRISH, (2015). "ENVIRONMENT CHANGE PREDICTION TO ADAPT CLIMATE-SMART AGRICULTURE USING BIG DATA ANALYTICS" [4]

The main aim of the paper is to predict changes in weather and help farmers in making agriculture-related decisions based on those changes. The paper has proposed a model to find solutions to modern world problems, such as worldwide food insecurity induced by frequent climate change, to predicting the impact of extreme weather events and mitigating its effect on global finance. They have made use of Big Data Analytics techniques to make an automatic prediction system. This paper builds the model based on the Hadoop framework.

First, they collected data from various sources like social media, sensor data, weather forecasts etc. and loaded the pre-processed data into HDFS. HDFS stores datasets and provides backup features. They focused on three factors while collecting data, namely, precipitation, temperature and cloud cover for the state of Karnataka. The authors have mainly used Hive for reading and processing data. Hive's strong SQL skills make it possible to process huge volumes of data stored in HDFS. Hive converts SQL queries into a series of MapReduce jobs. They have used MapReduce to analyse the data and as an execution engine suitable for large data processing and to improve the response speed for returning query results.

Then they implemented a prediction function for establishing forecast data through the k-means cluster algorithm. They used Apache Mahout to implement a logistic regression algorithm to predict the future based on the past data. For this, testing and training of data are performed. Then they evaluated the accuracy of the predicted result and represented the output using visualizations making use of the Flotend tool. Flot is a JavaScript plotting library. They used Pig script to perform analysis and the output was provided as an input to Flotend. They made various plots showing yearly and monthly average temperature for a particular region, maximum and minimum temperature, precipitation etc. The authors of the paper aim to improve their model such that it can be used for providing alerts in natural hazards in the future.

E. A. K. KUSHWAHA, S. BHATTACHARYA, "CROP YIELD PREDICTION USING AGRO ALGORITHM IN HADOOP" [5]

This paper aims to predict the crop yield and suggest crops based on it which would, in turn, increase the profit of the farmers and overall, the entire agriculture sector. It also focuses on improving the quality of the crops using datasets for diseases. They have used a new algorithm called Agro Algorithm to predict the crop yield and suggest crops [7] based on the crop yield and taking the soil type into consideration.

In section I of the paper, the authors have used weather datasets containing information about temperature, rainfall in mm, wind speed, evaporation, humidity etc. They further used the weather datasets to determine the type of soil. They also used datasets for crop diseases to determine the ideal weather conditions which would be suitable for a particular crop to grow. Section II presents the numerous methods that already exist for crop prediction and their drawbacks. Here they discussed techniques like clustering, soft computing techniques such as k-means and artificial neural networks. In section III of the paper, the paper talks about some basic knowledge that is required to improve the quality of crops, such as selection of plant and soil factors such as pH, which would play an important role in getting a good yield. The properties of the soil should also be known beforehand. It is also important to select the right seeds and estimate the right amount of fertiliser and pesticides required.

Section IV is the implementation. The implementation is performed on the Hadoop platform since the datasets are large. Normalisation is performed on the data stored in HDFS. This is done by taking the statistical average mean of data. First, the month in which the crop has to be sown is selected, then the classified data is used to predict the quality of soil and the recommended crop. The classification algorithm used is a simple statistical-based learning system. This prediction is represented using pie charts and this prediction is used to form five categories: ''very good'', ''good'', ''average'', ''bad'' and ''very bad''. Section V discusses the architecture used by the authors. They represented their architecture using a flowchart. Firstly, they collect multiple datasets related to agriculture and weather and perform the required analysis and classify the data. Then they used the classified data to predict the soil type and crop that can be sown.

In section VI of the paper, the authors talk about some issues and obstacles that are faced in quality farming in India [12]. These include technical gaps, small sizes of the farms, less availability of data and disparate harvesting systems.

In section VII, they conclude that the crop yield is improved by using weather, soil, crop and disease datasets. This in turn, boosts the standard of production of crops. It helps farmers immensely in selecting a crop suitable for the weather and soil type.

In the future, the authors aim to classify all the types of diseases for a particular crop and determine its cause which would further improve the quality of crops.

F. P. C. REDDY, DR. A. S. BABU, "SURVEY ON WEATHER PREDICTION USING BIG DATA ANALYTICS" [6]

In section I, this paper tells us about various methodologies to make weather predictions and discuss them in detail. Weather forecasts are important to prevent future damage, economic downfall and deaths, floods due to extreme

TABLE 1. Comparison of existing methods.

Method	Pros	Cons
Decision tree algorithm	It is easy to understand and explain.	It has high complexity and takes more time.
Multiple linear regression	It has the ability to identify outliers or anomalies.	Data is considered to be independent.
Backpropagation algorithm	It is simple and easy to program.	The actual performance on a particular problem is dependent on the input data.
K-means clustering	It is easy to implement.	Scaling with a number of dimensions.
Naïve Bayes Algorithm	Good for comparison in numerical variables.	It is an inadequate estimator.
Nearest Neighbour model	Can be used for both classification and regression.[16]	It does not work well with a large number of variables.
Artificial Neural Network	Good with non-linear data for large inputs.	Expensive computations and time-consuming.

weather conditions. Therefore, weather prediction such as rainfall and tropical cyclone prediction becomes very important. Weather prediction also proves helpful to farmers since it can prevent crop damage [19], [20]. The authors collected weather datasets that had data on humidity, rainfall, wind speed, temperature, air pressure, vapour pressure, sunlight intensity and various other factors. They collected this historical data from various sources. They have used big data analytics to study trends and patterns and predict short term as well as long term weather changes.

Section II describes the obstacles that occur in weather prediction. Conventional and statistical models do not give accurate results because the datasets are large and the output depends on assumptions [13]. It is hard to make a prediction model for the long term because the input parameters change rapidly, which are difficult to incorporate in an already built model. Noise in the dataset leads to inaccurate short-term predictions. In section III, they discuss the different methods adopted by researchers for weather prediction:

1) MapReduce

This model was used by a researcher for studying problems related to agricultural lands. They made a soil analysis system using various datasets like historical crop data, soil nutrient levels, fertilizers and manure used.

2) LINEAR REGRESSION AND MapReduce [22]

They collected data such as rainfall, temperature and humidity from weather stations and predicted weather, helping farmers in planting crops with good yield and cutting their costs.

$$f(q) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^2 \tag{1}$$

ξ denotes the difference between actual and predicted values and N is the number of numerical values in vector x in (1).

3) TIME DELAY RECURRENT NEURAL NETWORK AND FEED FORWARD NEURAL NETWORK [14]

They made use of evaporation, soil temperature and humidity, among other factors to predict rainfall. This provided daily, monthly as well as annual rainfall forecasting.

4) WAVELET ANN [15]

This method was used to predict daily average temperature and rainfall using various weather-related datasets.

$$e^{(\cos(\cos(-10.6\mathrm{EP}T_i))-(X_i/(\sin(\mathrm{RF}_1\mathrm{RH})^*\mathrm{RH}))+3.7292)} \tag{2}$$

where $X_i = \mathrm{EP}^\star e^{\log_{10}(4.3\mathrm{EP}+0.4941)} / (T_x + \mathrm{RH})$

In (2), RH means relative humidity, RF1 is 1-day previous rainfall and EP is evaporation.

TABLE 2. Review of existing work.

S. No.	Author Name	Data source	Parameters	Crop Studied	Methodology	Tools
1	R. Priya et. al [2]	Data collected from Krishi Vigyan Kendra and other satellite images, sensor recorded field data, irrigation related reports, crop data, weather data	Air temperature, relative humidity, wind speed, wind direction, soil temperature, soil moisture, radiation, diffusion rate and rainfall.	Rice, Maize, Cotton, Chillies	Naive Bayes, Map Reduce	HDFS, NB Classifier
2	R. M. G et. al [4]	Sensor data, weather forecasting, social media data and market trends.	Precipitation, temperature and cloud cover	Not specified	MapReduce, Logistic Regression	HDFS, Hive, Pig, Mahout, Flotend
3	W. Fan et. al [3]	Agricultural data (mainly weather data) in China, from 825 meteorological stations located in 34 districts. Weather data since 1951. Yield data was collected over the years.	Air pressure, temperature, relative humidity, precipitation, evaporation, wind speed, sunlight intensity and temperature at ground level.	Not specified	MapReduce, Nearest Neighbours, ARMA Model	HDFS
4	D. S. Zingade et. al [40]	IMD (Indian Meteorological Department)	Rainfall, temperature, soil and past year crop production	All possible crops	Multiple linear regression	Not specified
5	M. R. Bendre et. al [22]	KVR (Krishi Vidyapeeth Rahuri) Ahmednagar, India weather station from last 10 year, (1 Jan 2003 to 31 Dec 2013)	Daily minimum, maximum temperature, humidity and rainfall data	Not applicable	Map Reduce and Linear Regression	Hadoop, Google File System (GFS)
6	N. Gandhi et. al [41]	Publicly accessible records of the Indian Government for the years 1998 to 2002 (data of 27 districts of Maharashtra used)	Precipitation, minimum temperature, average temperature, maximum temperature, reference crop evapotranspiration, area, production and yield for the Kharif season	Rice	Sequential Minimal Optimisation (SMO) classifier	WEKA

TABLE 2. ***(Continued.)*** **Review of existing work.**

7	W. A. Goya et. al [42]	The CNCD (Canadian National Climate Data)	Temperature, Solar radiation, Evaporation, Wind speed, Rainfall, humidity	Not applicable	Map function Algorithms and Map Reduce job	Hadoop Distributed File System (HDFS) and Hadoop MapReduce.
8	S. Brdar et. al [43]	Data collected from 1999 to 2008 in the Serbian province of Vojvodina from the internal database of the Department of field and vegetable crops at the Faculty of Agriculture in Novi Sad.	Attributes used in this study are maximal (Tmax), minimal (Tmin) and average (Tavg) monthly air temperatures, as well as overall monthly hydrological cycle attributes: precipitation in mm (Pmm) and evapotranspiration	Maise, Soybean, Sugar beet	Support Vector Machines (SVM) regression	R package e1071
9	*Q. Huang et. al* [20]	European Centre for Medium-range Weather Forecasting (ECMWF), Remote sensing application center of Ministry of agriculture and Department of crop farming administration of the ministry of agriculture	Daily minimum and maximum temperature, rainfall, radiation, snow depth, vapour pressure and wind speed	Winter Wheat, Spring Wheat, Maize, Rice and Soybean	Regression and Scenario analysis.	China CGMS, JRCs "VIEWER" tool, CGMS Statistical Tool (CST), calibration platform (Calplat) tool
10	N. Gandhi et. al [44]	Publicly accessible records of the Indian Government for the years 1998 to 2002 (data of 27 districts of Maharashtra used)	Precipitation, minimum temperature, average temperature, maximum temperature, reference crop evapotranspiration and yield for Kharif Season	Rice	Artificial Neural Networks using backpropagation technique with Multilayer Perceptron	WEKA
11	Proposed Literature	Kaggle, University website and data found manually from websites (from 1951-2010)	Rainfall, Temperature, Humidity, Windspeed, Soil, Seed	125 Crops	MapReduce Framework, K-Means Clustering	Hadoop, NetBeans, Python, Flask

These comparisons between different methodologies help in finding the best-suited one for a particular prediction. Section IV concludes that MapReduce and Linear Regression models are better than other techniques to perform weather and climate forecasting as these methods provide an accurate result.

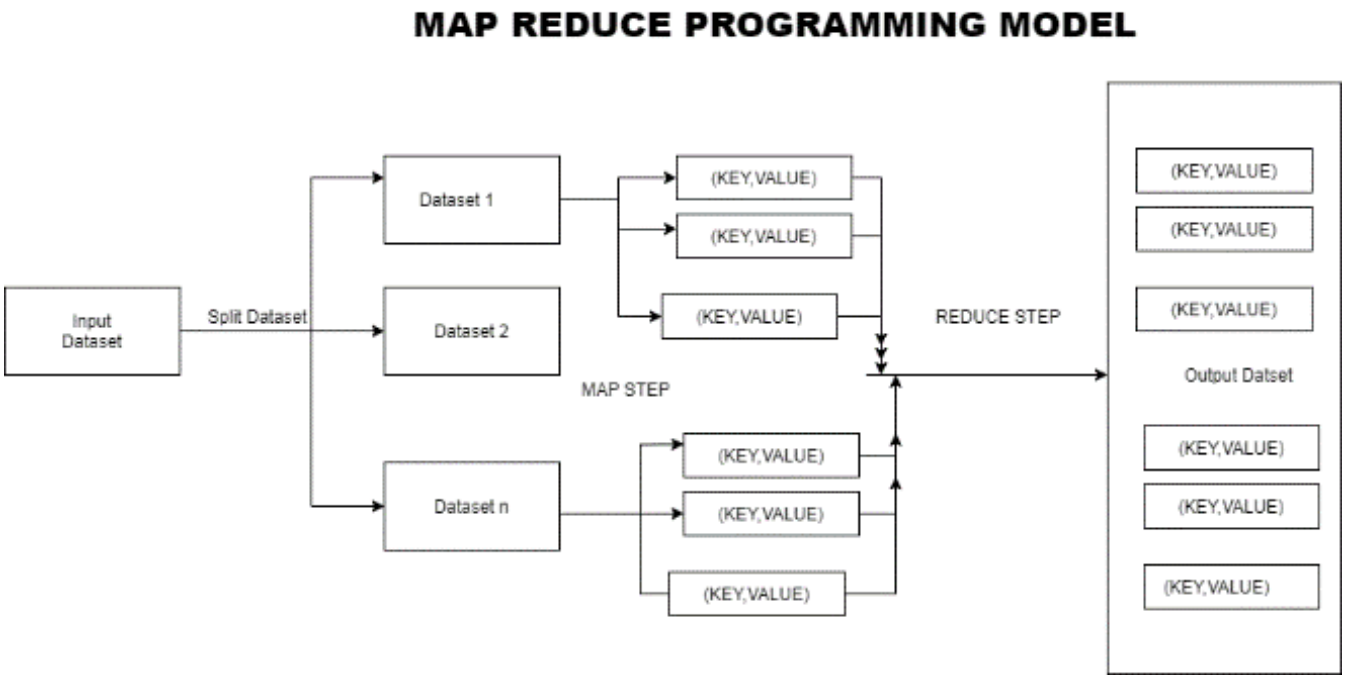


FIGURE 1. Proposed approach and map reduce programming model.

It was concluded from this literature review that MapReduce should be used as an efficient programming model for computing very large datasets of weather and climate with ease and high performance. Further, we concluded that using K means clustering on our final generated datasets would help us identify the relation between the crop and produce per area of the particular region.

III. PROPOSED METHODOLOGY

After studying the previous work done, the main aim would be to process the data using MapReduce and frame a recommender algorithm in Python to extract output according to the seasonal conditions and region followed by executing k-means clustering and finding the mean produce per area a group of crops will give in a particular region.

Keeping in mind the previous work done in other papers, we have taken temperature, rainfall, wind speed, humidity, soil type and seed type as the deciding parameters of our system. Firstly, the raw data will be collected and pre-processed in a Python environment. Then this pre-processed data is used as an input for the MapReduce framework of Hadoop to process the data. MapReduce is a programming model for processing large amounts of data with a parallel, distributed algorithm [29]. This model is implemented on the collected datasets for faster processing. In this work, each dataset will be processed differently. In the MapReduce model, the dataset will be divided into key and column pairs as shown in Fig. 1 where the different parameters will be individually taken to perform a MapReduce. The year and region will be stored in the key, and the respective parameter for all the months will be taken as the value for the Map Function. In the Reduce function, these parameters will be calculated and assigned to crop seasons. For the Map function of the crop dataset, the region, year, season and crop will be assigned as the key and the produce and area will be taken as the value. Then the Reduce function will calculate the produce per area for each row of data where the region, year, season and crop will form the key and produce per area will become the value.

Next, we propose to combine all the map reduced datasets of the different parameters to form one final/super dataset and make a recommendation algorithm. Three parameters can be taken as user input: the month, region and state. Next, we will initialise the different agricultural seasons. Then depending upon the user input of the month, we assign it the respective season/s. For example, if the user input is November, then we could assign it Rabi/ Winter crops. Next, we parse through the data and select the three crops that give the best yield in that particular season and also the crops that give the best yield throughout the whole year in that particular state and region. These would be in two different data frames, one for the particular season and one for the whole year. Along with this, we output the temperature, rainfall, wind speed and humidity in which the crop had previously given the same output. We also mention the seed type to be used for each kind of soil for the crops in different regions as the availability of seed and soil type varies from region to region. The output of this recommendation function will be displayed on a Graphical user interface (website) designed on Flask using Python, where the user could input the required data and get the output from the system.

Next, we will use a K-means clustering model. Firstly, we will make an elbow graph to calculate the number of clusters/optimal value of K that is required. We will be utilising the Scikit-learn library for this purpose. Then we will be using the fit predict method to get the values of clusters. This will be done in the form of an array where numbers starting from 0 will represent the values of one cluster. Then the clusters will be plotted using the scatter method of the Matplotlib library. Each cluster centroid will be shown which would represent the average value of a cluster about which each crop would be plotted, and every cluster would be represented by a different colour.

To study the relationship between the produce per area, crops and the respective parameters, we would create several 3D graphs and scatter plots. The produce per area could be taken as the Y-axis and crops as the X-axis and we could change the Z-axis according to the parameters taken (here temperature, rainfall, humidity and wind speed). We will also study the relationship between soil and seed type using 2D bar graphs and scatter plots using the Matplotlib, seaborn and mpl_toolkits in Python by taking soil as the X-axis and the number of crops it supports growth for in Y-axis. Also, a graph could be made by taking Crops as the X-axis and Produce per area on the Y-axis to study which crop gives the maximum produce per area in the particular region.

IV. IMPLEMENTATION

A. DATA COLLECTION

Various data sets were collected during this step. Facing a little difficulty, we found seven datasets related to our workflow and need from Kaggle [27] and a university website [28].

Name of datasets:

1. crop_production.csv
2. INBOMBAY.csv
3. INCHENAI.csv
4. INCALCUT.csv
5. INDELHI.csv
6. portblair.csv
7. rainfall in India 1901-2015.csv

The crop_production.csv contained all the states and their districts. It contains data of 125 crops along with their production of each crop and area it was sown in from the year 2000 to 2014 for six seasons: Kharif, Rabi, Summer, Winter, Autumn and the whole year. INBOMBAY.csv, INCHENAI.csv, INCALCUT.csv, INDELHI.csv, portblair.csv contain the daily average temperature of the cities from the year 1995 to 2020. Rainfall in India 1901-2015.csv has the average monthly rainfall of different subdivisions across the country from the year 1901 to 2015. The soil, seed, humidity and wind speed data were manually collected from various websites like agricoop [36], Department of agriculture government of maharashtra [37], weatheronline [38], and climate-org [39].

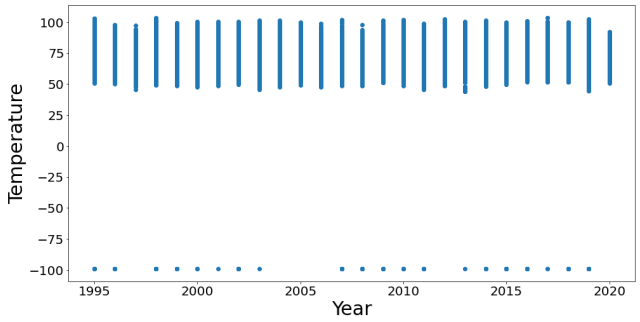


FIGURE 2. Graph before applying InterQuartile range.

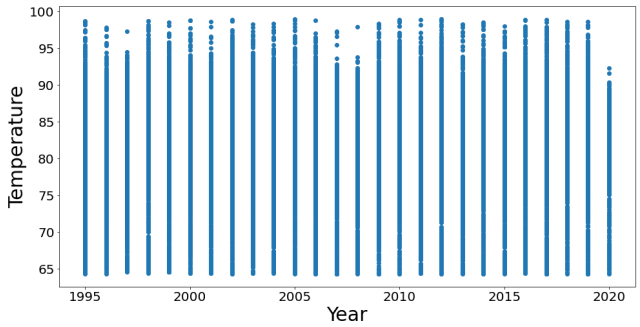


FIGURE 3. Applying InterQuartile range.

B. PRE-PROCESSING DATA

Here, the collected datasets were combined and cleaned. We uploaded our datasets on the Colab notebook and used pandas data frame to drop the useless columns and retain the ones important to us. We used NumPy, SciPy libraries for our calculations. A few index columns were added for future calculations. Interpolation, as represented in (3), was used to find the estimated value for the missing value in the dataset statistically [45], [46], [32]. In our datasets, we have established the value of certain numerical columns which had NA values, such as the month columns for the rainfall dataset.

$$\text{dataframe\_name.interpolate}\ (\text{method} = \text{`linear'}, \text{direction} = \text{`forward'}, \text{inplace} = \text{True}) \tag{3}$$

where dataframe_name is the name of the dataset in use and the interpolation is done in the forward direction linearly. In the next step, the redundant and dirty data was eliminated using the IQR and z-score method for detecting and deleting the outliers and interpolated the data in a few datasets where it was deemed fit. The Interquartile Range Method as in (4.1-4.4) was used to remove the outliers in the temperature datasets. The 25th and 75th percentile are found, and then 1.5 is taken as the factor because considering only three deviations is appropriate, i.e., 1.5 multiplied by 2, as anything greater than or less than these deviations on the sides above 75th and below 25th percentile will give

```
SUBDIVISION,YEAR,JAN,FEB,MAR,APR,MAY,JUN,JUL,AUG,SEP,OCT,NOV,DEC,ANNUAL,Jan-Feb,Mar-May,Jun-Sep,Oct-Dec
ANDAMAN & NICOBAR ISLANDS,1901,49.2,87.1,29.2,2.3,528.8,517.5,365.1,481.1,332.6,388.5,558.2,33.6,3373.2,136.3,560.3,1696.3,980.3
ANDAMAN & NICOBAR ISLANDS,1902,0,159.8,12.2,0,446.1,537.1,228.9,753.7,666.2,197.2,359,160.5,3520.7,159.8,458.3,2185.9,716.7
ANDAMAN & NICOBAR ISLANDS,1903,12.7,144,0,1,235.1,479.9,728.4,326.7,339,181.2,284.4,225,2957.4,156.7,236.1,1874,690.6
ANDAMAN & NICOBAR ISLANDS,1904,9.4,14.7,0,202.4,304.5,495.1,502,160.1,820.4,222.2,308.7,40.1,3079.6,24.1,506.9,1977.6,571
ANDAMAN & NICOBAR ISLANDS,1905,1.3,0,3.3,26.9,279.5,628.7,368.7,330.5,297,260.7,25.4,344.7,2566.7,1.3,309.7,1624.9,630.8
ANDAMAN & NICOBAR ISLANDS,1906,36.6,0,0,0,556.1,733.3,247.7,320.5,164.3,267.8,128.9,79.2,2534.4,36.6,556.1,1465.8,475.9
ANDAMAN & NICOBAR ISLANDS,1907,110.7,0,113.3,21.6,616.3,305.2,443.9,377.6,200.4,264.4,648.9,245.6,3347.9,110.7,751.2,1327.1,1158.9
.
.
.
.
LAKSHADWEEP,2009,4.7,1.5,0.1,18.1,162.1,401.2,266.4,185,145.1,87.4,166.2,132.3,1570.1,6.2,180.3,997.7,385.9
LAKSHADWEEP,2010,18.8,0,1.2,35.6,79,318.9,336.7,335.1,161.5,155.4,201.5,81.5,1725.2,18.8,115.8,1152.2,438.4
LAKSHADWEEP,2011,5.1,2.8,3.1,85.9,107.2,153.6,350.2,254,255.2,117.4,184.3,14.9,1533.7,7.9,196.2,1013,316.6
LAKSHADWEEP,2012,19.2,0.1,1.6,76.8,21.2,327,231.5,381.2,179.8,145.9,12.4,8.8,1405.5,19.3,99.6,1119.5,167.1
LAKSHADWEEP,2013,26.2,34.4,37.5,5.3,88.3,426.2,296.4,154.4,180,72.8,78.1,26.7,1426.3,60.6,131.1,1057,177.6
LAKSHADWEEP,2014,53.2,16.1,4.4,14.9,57.4,244.1,116.1,466.1,132.2,169.2,59,62.3,1395,69.3,76.7,958.5,290.5
LAKSHADWEEP,2015,2.2,0.5,3.7,87.1,133.1,296.6,257.5,146.4,160.4,165.4,231,159,1642.9,2.7,223.9,860.9,555.4
```

FIGURE 4. Input of map reduce.

```
1901, ASSAM & MEGHALAYA 77.2,1711.0,460.6,46.6,279.3,2498.7,115.6,1.2
1901, BIHAR 200.09999,786.69995,78.6,72.6,15.6,952.3999,7.3,0.1
1901, CHHATTISGARH 200.6,1142.1001,51.699997,165.5,27.699999,1386.9,0.4,0.0
1901, COASTAL ANDHRA PRADESH 107.299995,449.7,104.6,99.7,338.2,993.7,164.8,1.5
1901, EAST MADHYA PRADESH 268.6,1207.6,32.6,88.1,5.9,1332.7,0.0,0.0
1901, EAST RAJASTHAN 33.4,362.90002,8.6,30.5,9.8,412.59998,0.0,0.8
1901, EAST UTTAR PRADESH 102.899994,749.30005,22.900002,94.7,5.0,873.19995,0.1,2.1
.
.
.
2015,SUB HIMALAYAN WEST BENGAL & SIKKIM 238.0,1883.0,518.4,35.1,77.399994,2518.5002,23.8,9.0
2015,TELANGANA 93.3,690.1,132.0,26.5,15.900001,857.19995,0.3,1.7
2015,UTTARAKHAND 246.4,881.49994,222.6,118.8,19.199999,1247.6,2.4,7.2
2015,VIDARBHA 106.90001,848.19995,107.200005,38.2,7.0,993.60004,0.0,0.2
2015,WEST MADHYA PRADESH 100.3,914.5,68.8,46.800003,11.3,1042.2001,0.3,1.0
2015,WEST RAJASTHAN 33.899998,384.00003,71.0,3.3000002,1.2,458.50003,0.1,0.0
2015,WEST UTTAR PRADESH 105.700005,436.2,95.9,38.8,8.9,582.80005,2.0,3.0
```

FIGURE 5. Output of map reduce.

inaccurate results [30].

$$Q1 = \text{dataframe\_name[column\_name].quantile(0.25)} \tag{4.1}$$

$$Q3 = \text{dataframe\_name[column\_name].quantile(0.75)} \tag{4.2}$$

$$IQR = Q3 - Q1 \tag{4.3}$$

$$\begin{aligned} \text{output} = {} & \text{dataframe\_name} \\ & \times[\sim ((\text{dataframe\_name[column\_name]} \\ & < (Q1 - 1.5^{*}IQR)) \\ & |(\text{dataframe\_name[column\_name]} \\ & > (Q3 + 1.5^{*}IQR))) \end{aligned} \tag{4.4}$$

where, *Q1* and *Q3* are the first 25th and last 75th range and *IQR* is the interquartile range. *dataframe_name* is the name of the dataset in use and *column_name* is the column we want to apply it to. In Fig. 2 the outliers before cleaning the dataset are plotted and in Fig. 3, the cleaned dataset with no outliers is plotted using matplotlib.

Z score as shown in (5) is also used to remove the outliers here. Z score is also used to remove the outliers here. We take 3 as a factor here because anything above a positive 3 and below negative three will bring us inaccurate results and hence are outliers [31].

$$\begin{aligned} & \text{constrains} \\ & = \text{dataframe\_name.select\_dtypes(include=[numpy.number])} \\ & \quad .\text{apply(lambdax : numpy.abs(scipy.stats.zscore(x))} < 3) \\ & \quad .\text{all(axis} = 1)\text{dataframe\_name.drop(dataframe\_name} \\ & \quad .\text{index}[\sim \text{constrains}], \text{inplace} = \text{True}) \end{aligned} \tag{5}$$

where, *dataframe_name* is the name of the dataset in use and x is the data we are performing z-score on.

C. MAP REDUCE

The MapReduce model was implemented on the cleaned data, where the dataset is divided into key and column pairs. For the INBOMBAY.csv, INCHENAI.csv, INCALCUT.csv and INDELHI.csv, in the Map Function first, we took the month, year and region as the key and the temperatures are taken as the value. Then in the Reduce Function calculation was performed to find the average monthly temperature of the regions and store the month, year and region as the key and the average temperature as the value. The year and region will be stored in the key and the respective parameter will

```
state=input("Enter state:")
district=input("Enter district")
month=input("Enter Month")
# state="Maharashtra"
# district="AHMEDNAGAR"
# month="November"
season=""
season1=""
season2=""
season4=""
if(month=="January"or month=="December" or month=="November" or month=="February"or month=="March"):
  season='Rabi     '
  # print(season)
if(month=="October" or month=="November"):
  season1='Autumn     '
  # print(season1)
if(month=="June" or month=="July" or month=="August" or month=="September"):
  season2='Kharif     '
  # print(season2)
if (month=="March" or month=="April" or month=="May"):
  season4='Summer     '
season3='Whole Year '
dp=crop.loc[(crop['State'] == state) & (crop['District'] == district)& ((crop['Season']==season) |
                                                                        (crop['Season']==season1) |
                                                                        (crop['Season']==season2) |
                                                                        (crop['Season']==season4))]
# print(dp)
wy=crop.loc[(crop['State'] == state) & (crop['District'] == district)& ((crop['Season']==season3))]
# print(wy)
```

FIGURE 6. Taking user input.

```
str1=","
print("The following crops could be sown in the given order according to your input month :"+str1.join(crop_list)+"\n")
if(len(crop_list)<3):
  t=len(crop_list)
else:
  t=3
if(len(crop_list1)<3):
 t1=len(crop_list1)
else:
 t1=3
print("However highly recommended would be as follows:")
print()
for i in range(t):
  print(crop_list[i]+" could be sown if the temperature is approximately: ",temp_list[i]," and the observed rainfall is:
   ",rain_list[i],"for a expected produce per area of",prod_list[i],".")
  print(".Variety should be",seed1_list[i],"if soil type is",s1_list[i],"and",seed2_list[i],"if soil is",s2_list[i],".\n")
  print("The best produce could be expected if the humidity", hum_list[i]," is and the windspeed is", wind_list[i])
print()
print()
print("The crops metioned below can give a good produce theroughout the year: \n")
for i in range(t1):
  print(crop_list1[i]+" could be sown if the temperature is approximately: ",temp_list1[i]," and the observed rainfall is:
  ",rain_list1[i],"for a expected produce per area of",prod_list1[i],".")
  print(".Variety should be",seed1_list1[i],"if soil type is",s1_list1[i],"and",seed2_list1[i],"if soil is",s2_list1[i],".\n")
  print("The best produce could be expected if the humidity", hum_list1[i]," is and the windspeed is", wind_list1[i])
```

FIGURE 7. Output of recommender function.

be taken as the value month wise for the Map Function. In the Reduce function, these parameters will be calculated and assigned to crop seasons like Rabi, Kharif, Autumn, Winter, Summer, Whole Year, and the temporary December and November temperatures for the next year Rabi and Winter calculations. For the rainfall in India 1901-2015.csv in Fig. 4, MapReduce was performed to calculate cumulative rainfall in the different agricultural seasons according to the dataset. The output for the same is displayed in Fig. 5. For the Map function of the crop dataset, the region, year, season and crop will be assigned as the key and the produce and area will be taken as the value. Then the Reduce function will calculate the produce per area for each row of data where the region, year, season and crop will form the key and produce per area will become the value. The produce per area of all the crops was found from the crop_production.csv dataset, which will suggest a particular crop according to the region.

D. PYTHON RECOMMENDATION FUNCTION

The cleaned and map reduced datasets of rainfall, temperature and crop production were combined along with the manually collected wind speed, humidity, soil type and seed type data to form one final dataset. We took the input to get the state, region and month from the user themselves as depicted in Fig. 6. Next, on the basis of the month input by the user,

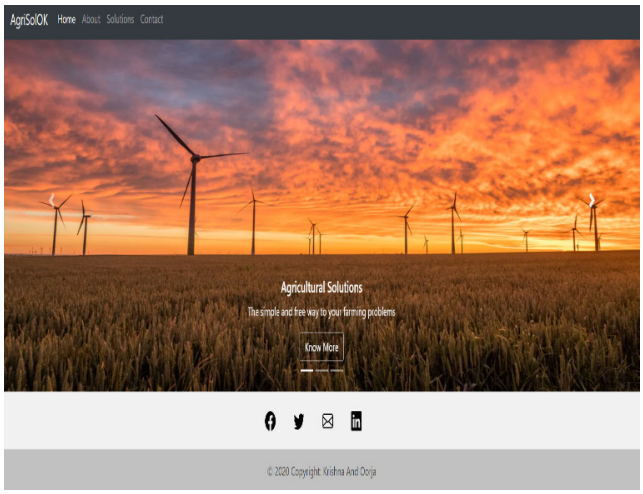


FIGURE 8. Landing page.

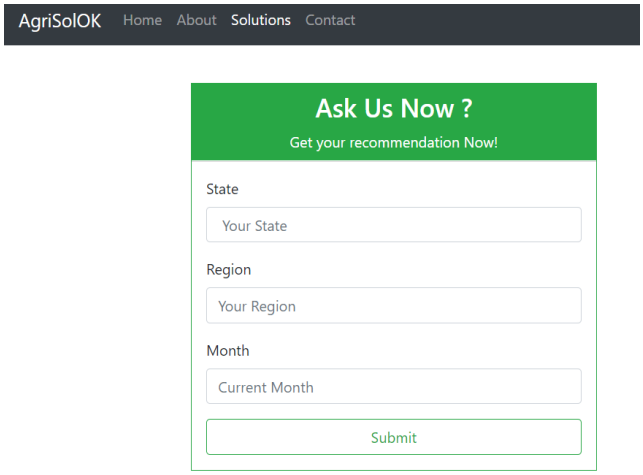


FIGURE 9. Solutions page.

we assigned the agricultural season/s. After this, depending upon the season assigned, we have parsed through our data to collect the best yield of three crops in the selected input region and state. Besides this, we have extracted the top three crops giving the best yield throughout the year. Then depending upon the crop, the seed type and soil type with two different varieties are suggested as availability of the soil and the varieties of suitable and available seed types vary from region to region. Further, the temperature, rainfall, wind speed and humidity of the region at which the crop had given such stellar outputs is also displayed for reference.

This function runs the provided input through the above algorithm to give a favourable recommender output as shown in Fig. 7. We have taken months of Rabi, Kharif, Summer, Winter, Autumn and Whole year into consideration.

E. WEBSITE

A minimalistic website is designed as a graphical user interface for the user. The front end of the website has been made on Flask 2.0.0 in a virtual environment using Python 3,

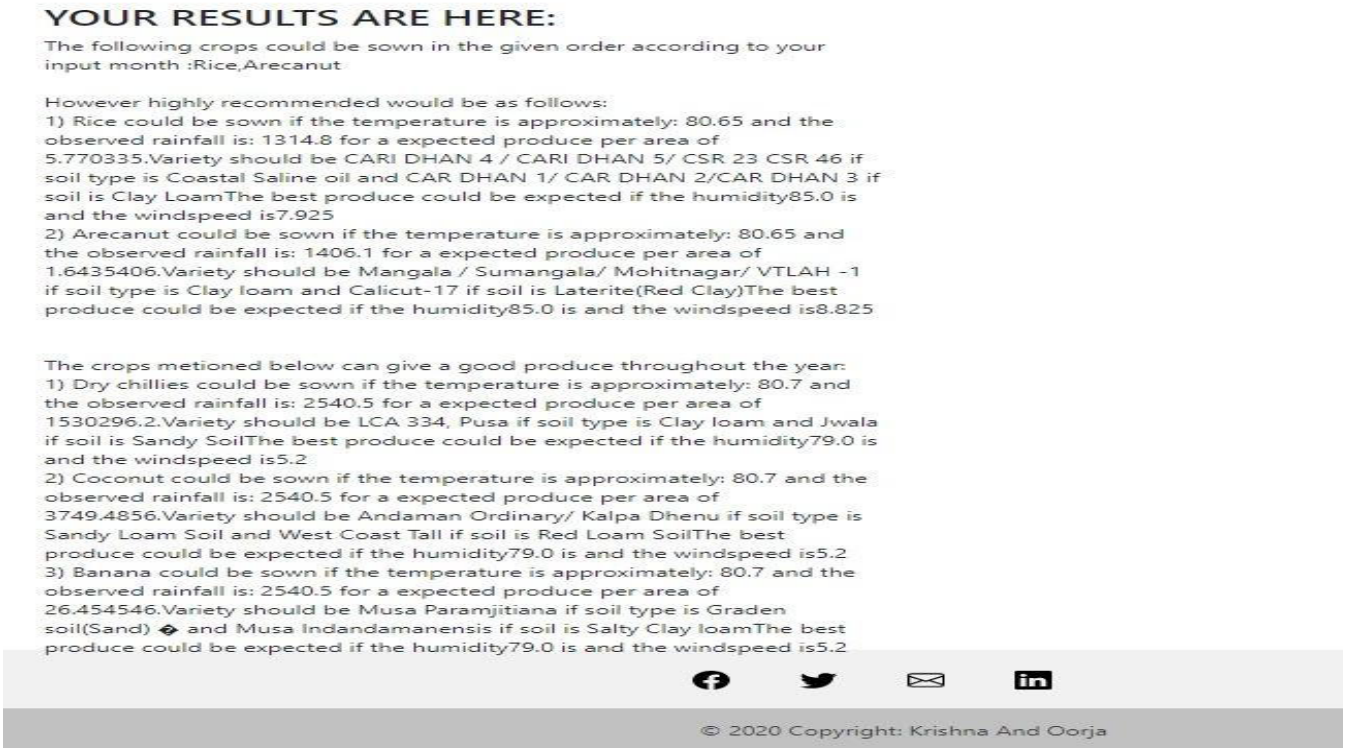


FIGURE 10. Shows the output of the recommendation.

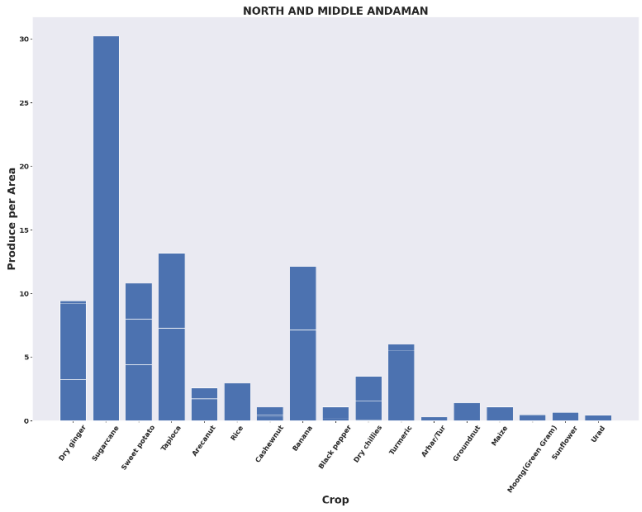


FIGURE 11. Bar graph of total production per area to crop.

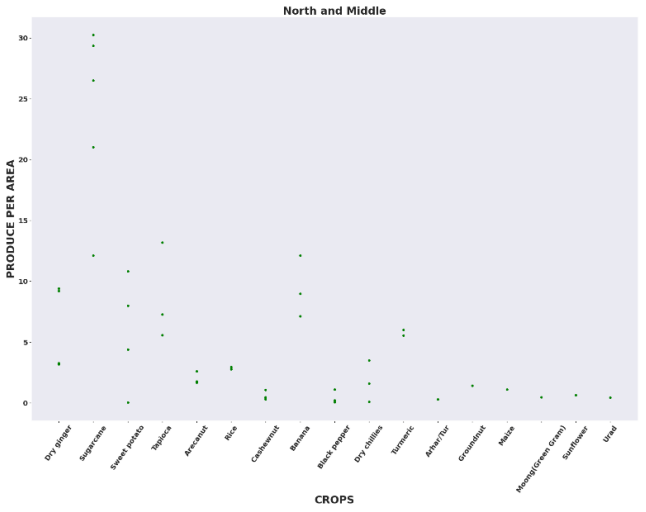


FIGURE 12. Scatter plot of production per area to crop.

HTML 5.0, CSS3, Bootstrap 4 and Jinja2 templates. The website has four pages: Home, About, Contact, Solutions. The Home page, as shown in Fig. 8, is a landing page to the website. The About page gives some brief information on what our website is and how to access different components within the website. Next is the Contact page form, where the users can send us messages or requests if required. And last but most important is the Solutions page, which is shown in Fig. 9. It takes the desired input from the user and shows them the predicted output, that is, the top three seasonal crops with 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 that desired output. It also suggests two kinds of suitable soil and the respective seeds where the crop can be grown as shown in Fig. 10.

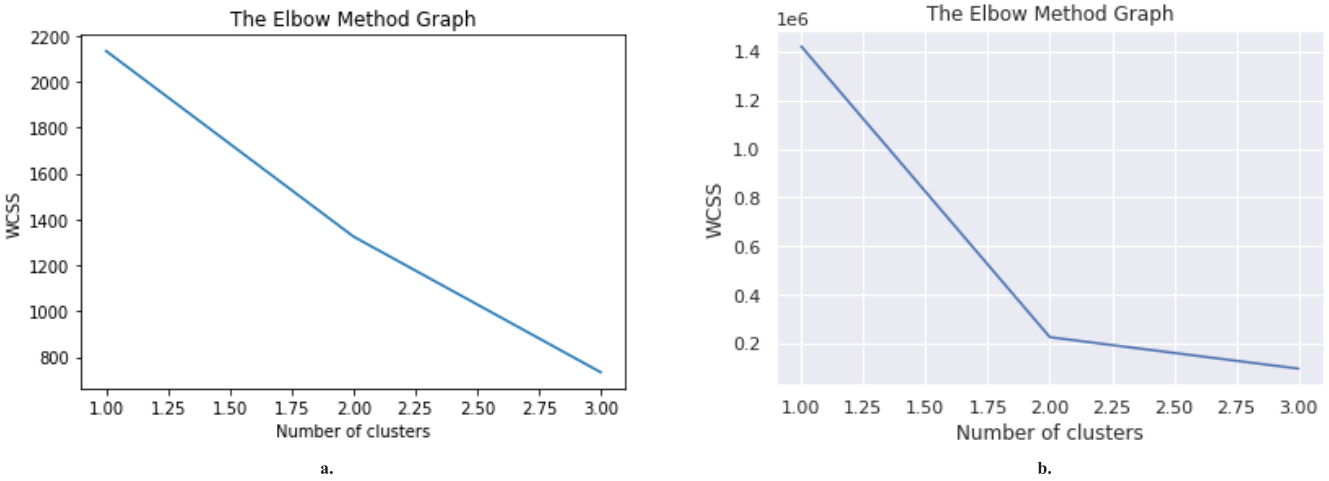


FIGURE 13. (a) Elbow graph to find the number of clusters for Nicobar region. (b) Elbow graph to find number of clusters for Ahmednagar, Maharashtra.

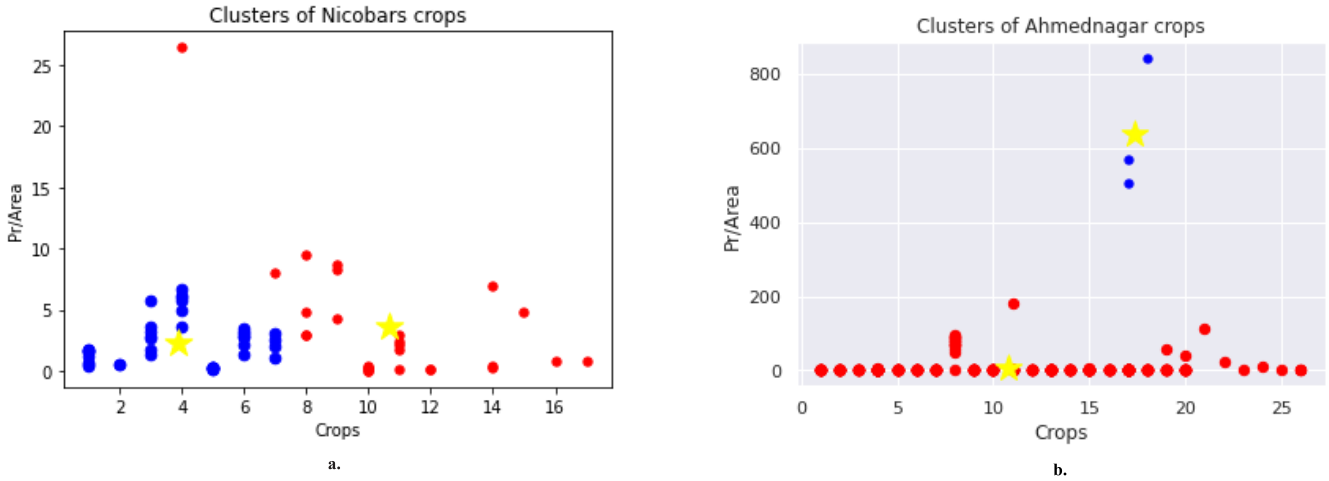


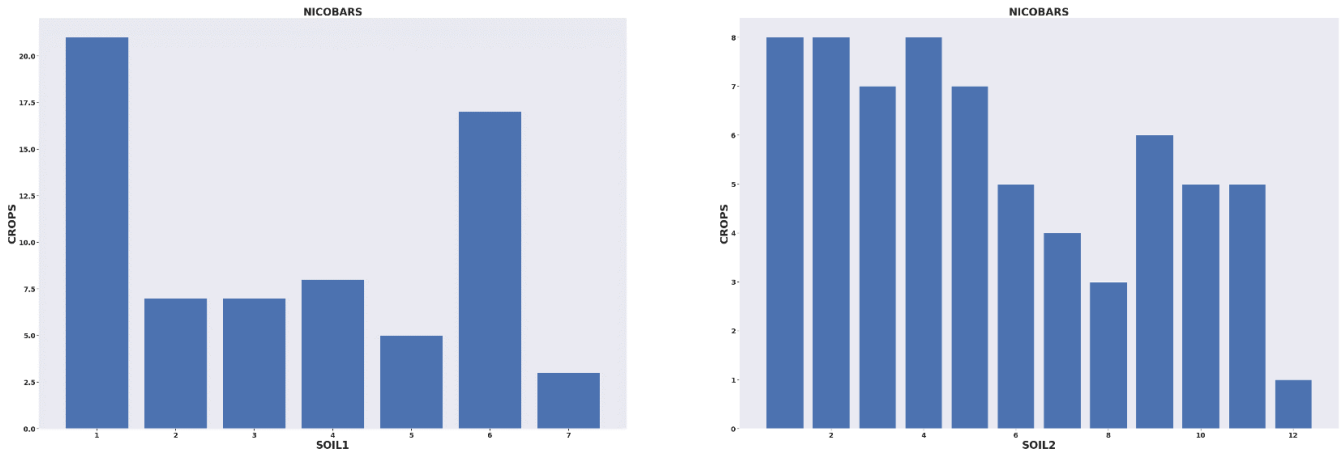
FIGURE 14. (a) Output of k-means clustering (Nicobar region). (b) Output of k-means clustering (Ahmednagar, Maharashtra).

F. VISUALISATION

The super dataset is used to find the relation between produce per area and a particular crop using a bar graph and scatter plot for a particular region. It can be seen in the bar graph represented in Fig. 11 that sugarcane has the highest production per area in North and Middle Andaman Region. In Fig. 12, the scatter plot of the same region is made to plot the initial produce per area and the crops before clustering to differentiate before and after the k-means clustering algorithm.

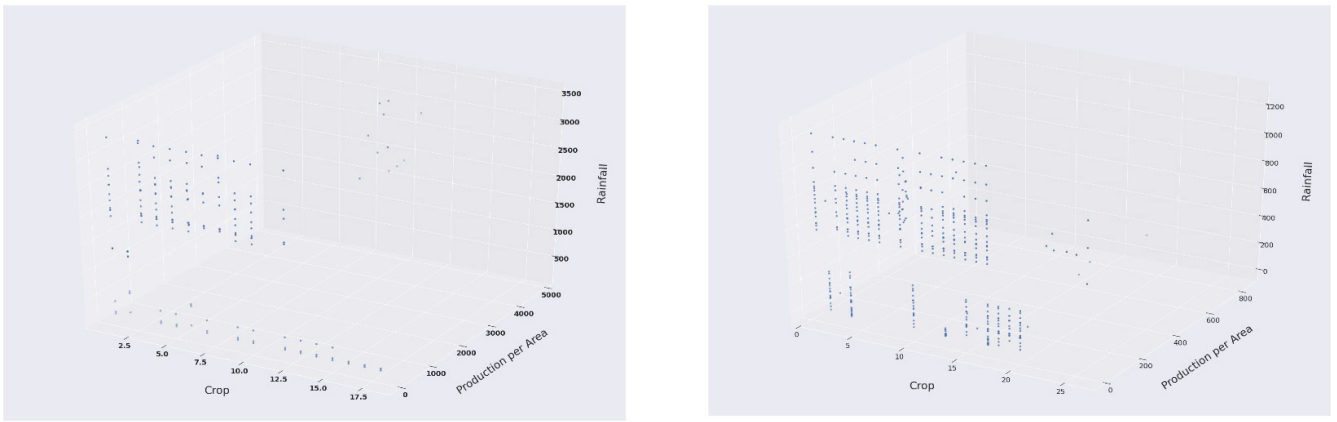
The elbow graph has been plotted to find the number of clusters that should be made for a particular region's crops according to its produce per area as shown in Fig. 13 a and b. The point where the graph bends give the number of clusters the dataset should ideally have. The elbow of the graph was found to be 2 for both Nicobars as well as Ahmednagar, Maharashtra as can be seen in the figures. The clustering algorithm is then applied and the clusters are formed with the crops plotting their produce per area around the cluster centroids. From the graph in Fig. 14 a, we can deduce that the mean produce per area of the first ten crops in the Nicobar is around 2.5 units and the mean produce per area of the next ten crops in the graph is around 4 units. In Ahmednagar, however, the clusters are formed in such a way that most crops have a produce per area less than 1 unit per area and hence have a cluster center of 0, as shown in Fig. 14 b. While the other cluster is of crops giving very high output per Area, giving an average of about 600 units. These are usually oilseeds grown by farmers.

Next, bar graphs have been plotted to check the relationship between a soil and the different types of seeds that can be sown in it. The graphs in Fig. 15 a and b show 8 different types of soil found in the Nicobar region and the number of varieties of seed that can be sown in the same. Each soil type has been assigned a number from 1 to 7, and the number of crops that can be planted in a particular soil type is calculated and shown on the y-axis. It can be concluded from the graph that Clay



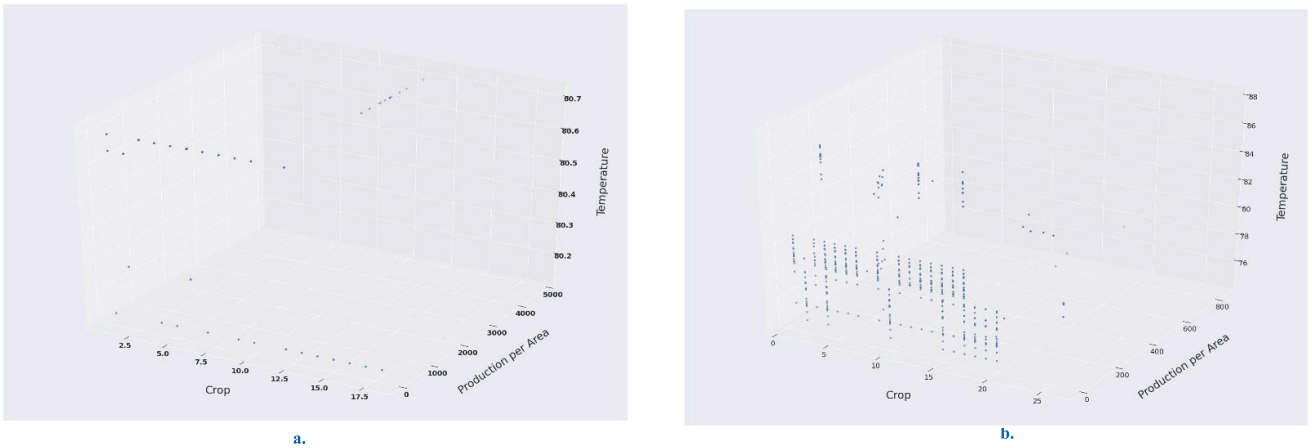
a. b.

FIGURE 15. (a) Graph showing the relation between soil type and seeds grown in Nicobar region. (b) Graph showing the relation between soil type and seeds grown in that soil in Nicobar region.



a. b.

FIGURE 16. (a) 3D scatter plot showing the relationship between crop, production per area and rainfall for Andaman and Nicobar Islands. (b) 3D scatter plot showing the relationship between crop, production per area and rainfall for Ahmednagar, Maharashtra.



a. b.

FIGURE 17. (a) 3D scatter plot showing the relationship between crop, production per area and temperature for Andaman and Nicobar Islands. (b) 3D scatter plot showing the relationship between crop, production per area and temperature for Ahmednagar, Maharashtra.

Loam Soil holds the capacity to sow 17 crops with different seed varieties. In Fig.15 b, Laterite (Red Clay), Clay Loam and Sandy Loam show the highest capacity to sow crops.

Further, we have depicted 3D plots of the Andaman and Nicobar Islands and the Ahmednagar region of Maharashtra. The plots in Fig.16 a and b represent the relationship between

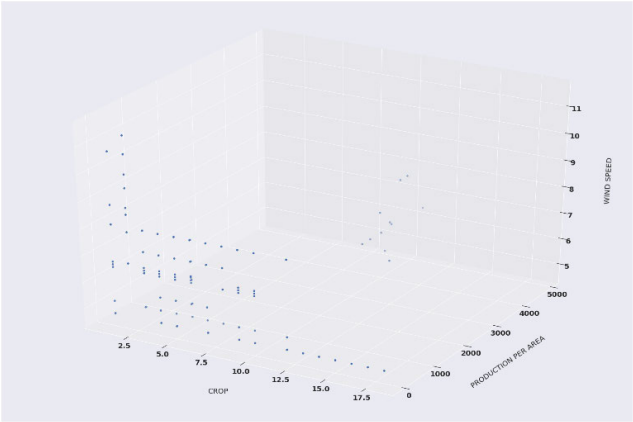


FIGURE 18. 3D scatter plot showing the relationship between crop, production per area and wind speed for Andaman and Nicobar Island.

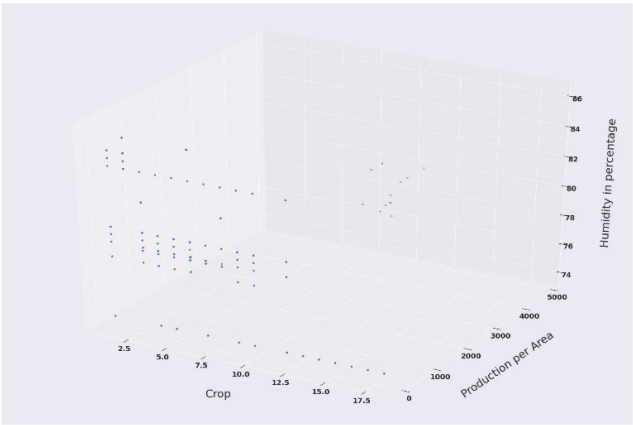


FIGURE 19. 3D scatter plot showing the relationship between crop, production per area and humidity for Andaman and Nicobar Islands.

the crops, production per area and rainfall. Fig. 16 a shows the effect of rainfall on Andaman and Nicobar Islands whereas Fig. 16 b shows the effect on Ahmednagar, Maharashtra. In the case of Andaman and Nicobar Islands, a high produce per area could be expected if the rainfall is around 500-2000mm throughout the year. For Ahmednagar, it can be seen that most of the crops give a high produce per area if the rainfall is from 0 to 800mm every year.

Fig. 17 a and b show the relationship between various crops, their production per area and the temperature they were grown in for Andaman and Nicobar and Ahmednagar respectively. In Andaman and Nicobar region, most of the crops give a high produce per area if the temperature is between 80.3 and 80.7, while for Ahmednagar, a high produce per area is obtained when the temperature is between 74 and 80 units.

When it comes to the relation between humidity, crop and production per area it can be noticed from Fig. 18 that most crops thrive in 74 %-78% humidity, whereas a few crops require higher humidity between 80%-84%. The scatter plot of Fig. 19 depicting the relation of wind speed with crop and production per area concludes that most crops require a wind speed of 5-8 units to grow the best, whereas some of the crops can be seen thriving irrespective of the increase or decrease in wind speed.

All the 3D graphs plotted above show how the types of crops and their production per area is affected by different factors like rainfall, temperature, wind speed, and humidity. It can be noted that all these parameters combine together to form a suitable environment to give a good produce per area for a major variety of crop.

V. CONCLUSION AND FUTURE SCOPE

The proposed work introduces a crop recommendation system and uses MapReduce and K-means clustering, which gives efficient results in terms of computations. The model focuses on a wide range of crops and their produce per area along with the soil type and seed types depending on the varieties used in a particular region. From the visualisation graphs of K-Means clustering, we can find the mean produce for a group of crops. The algorithms that have been used for the recommender function and K-Means Clustering can be accessed on https://github.com/oorjagarg/WB-CPI. Also, the relation between parameters (like optimal temperature, seasonal rainfall, wind speed, humidity, soil availability, required seed types), crop and region has been studied and displayed using 2D and 3D graphs. The system is scalable and it can be used to find the recommended crops of other states in a similar manner as described in the methodology. This work can be further improved to eliminate the problem of disproportion in the production and requirement ratio if an aspect of humidity, wind speed can be added for all the regions and will give a more accurate recommendation. Factors like soil moisture, irrigation, cloud cover etc. may be included in the system to refine its output. Also, the recommender can be modified to warn about the diseases that can occur in a crop in a particular season and suggest the types of fertilizers or nutrients needed in the soil for the crop to grow and give its best yield.

ACKNOWLEDGMENT

This research was supported and funded from Universiti Tun Hussein Onn Malaysia under Industry Grant Vot No M029.

REFERENCES

[1] D. Bose. (2020). *Big Data Analytics in Agriculture*. [Online]. Available: https://www.researchgate.net/publication/339102917

[2] R. Priya, D. Ramesh, and E. Khosla, "Crop prediction on the region belts of india: A Naïve Bayes MapReduce precision agricultural model," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2018, pp. 99–104.

[3] W. Fan, C. Chong, G. Xiaoling, Y. Hua, and W. Juyun, "Prediction of crop yield using big data," in *Proc. 8th Int. Symp. Comput. Intell. Design (ISCID)*, Dec. 2015, pp. 255–260, doi: 10.1109/ISCID.2015.191.

[4] M. Ramya, C. Balaji, and L. Girish, "Environment change prediction to adapt climate-smart agriculture using big data," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 4, no. 5, pp. 1995–2000, 2015.

[5] A. K. Kushwaha and S. Bhattachrya, "Crop yield prediction using agro algorithm in Hadoop," *Int. J. Comput. Sci. Inf. Technol. Secur.*, vol. 5, no. 2, pp. 271–274, 2015.

[6] P. C. Reddy and A. S. D. Babu, "Survey on weather prediction using big data analystics," in *Proc. 2nd IEEE Int. Conf. Electr., Comput. Commun. Technol.*, Feb. 2017, pp. 1–6.

[7] P. S. Vijayabaskar, R. Sreemathi, and E. Keertanaa, "Crop prediction using predictive analytics," in *Proc. Int. Conf. Comput. Power, Energy Inf. Commun. (ICCPEIC)*, Mar. 2017, pp. 370–373.

[8] S. Rajeswari, R. Scholar, K. Suthendran, and K. Rajakumar, "A smart agricultural model by integrating IoT, mobile and cloud-based big data analytics," in *Proc. Int. Conf. Intell. Comput. Control*, 2017, pp. 1–5.

[9] K. Charvat, K. C. Junior, T. Reznik, V. Lukas, K. Jedlicka, R. Palma, and R. Berzins, "Advanced visualisation of big data for agriculture as part of databio development," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 415–418.
[10] P. Shah, D. Hiremath, and S. Chaudhary, "Towards development of spark based agricultural information system including geo-spatial data," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 3476–3481.
[11] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5432–5435, Apr. 2009.
[12] H. Guo and H. L. Viktor, "Multirelational classification: A multiple view approach," *Knowl. Inf. Syst.*, vol. 17, no. 3, pp. 287–312, 2008, doi: 10.1007/s10115-008-0127-5.
[13] G. Geetha and R. S. Selvaraj, "Prediction of monthly rainfall in Chennai using back propagation neural network model," *Int. J. Eng. Sci. Technol.*, vol. 3, no. 1, 2011.
[14] M. P. Darji, V. K. Dabhi, and H. B. Prajapati, "Rainfall forecasting using neural network: A survey," in *Proc. Int. Conf. Adv. Comput. Eng. Appl.*, Mar. 2015, pp. 706–713, doi: 10.1109/ICACEA.2015.7164782.
[15] V. K. Dabhi and S. Chaudhary, "Hybrid wavelet-postfix-GP model for rainfall prediction of Anand region of India," *Adv. Artif. Intell.*, vol. 2014, pp. 1–11, Jun. 2014, doi: 10.1155/2014/717803.
[16] M. Motlhabi, P. Pantsi, and R. Netshiya, "A machine learning deep-dive analysis into network logs," in *Proc. ICCWS*, 2021, p. 213, doi: 10.34190/IWS.21.019.
[17] Z. Doshi, S. Nadkarni, R. Agrawal, and N. Shah, "AgroConsultant: Intelligent crop recommendation system using machine learning algorithms," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2018, pp. 1–6.
[18] M. Kumar and M. Nagar, "Big data analytics in agriculture and distribution channel," in *Proc. Int. Conf. Comput. Methodol. Commun. (ICCMC)*, Jul. 2017, pp. 384–387.
[19] A. Savla, N. Israni, P. Dhawan, A. Mandholia, H. Bhadada, and S. Bhardwaj, "Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture," in *Proc. Int. Conf. Innov. Inf., Embedded Commun. Syst. (ICIIECS)*, Mar. 2015, pp. 1–7.
[20] Q. Huang, Z. Chen, W. Wu, A. de Wit, F. Teng, and D. Li, "China crop growth monitoring system-methodology and operational activities overview," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 2961–2964.
[21] G. M. Alves and P. E. Cruvinel, "Big data environment for agricultural soil analysis from CT digital images," in *Proc. IEEE 10th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2016, pp. 429–431, doi: 10.1109/ICSC.2016.80.
[22] M. R. Bendre, R. C. Thool, and V. R. Thool, "Big data in precision agriculture: Weather forecasting for future farming," in *Proc. 1st Int. Conf. Next Gener. Comput. Technol. (NGCT)*, Sep. 2015, pp. 744–750, doi: 10.1109/NGCT.2015.7375220.
[23] J. Warner. (Accessed: Feb. 25, 2021). *Agriculture is the New Bee for Big Data Analytics!* Irishtechnews.ie. [Online]. Available: https://irishtechnews.ie/agriculture-is-the-new-bee-for-big-data-analytics/
[24] Zentut. (Accessed: Feb. 28, 2021). *Data Mining Techniques*. Zentut.com. [Online]. Available: https://www.zentut.com/data-mining/data-mining-techniques/
[25] M. A. Beyer and D. Laney, "The importance of 'big data': A definition," Gartner, Stamford, CT, USA, Tech. Rep. 1644565, 2012.
[26] X. E. Wang and W. L. Decker, "The use of distance coefficient in the research on agro climatological resemblance," *J. Nanjing Inst. Meteorol.*, vol. 12, no. 2, pp. 187–199, 1989.
[27] Kaggle. (Accessed: Mar. 5, 2021). *Datasets*. Kaggle.com. [Online]. Available: https://www.kaggle.com/datasets
[28] The University of Dayton. (Accessed: Mar. 8, 2021). *Average Daily Temperature Archive*. Udayton.edu. [Online]. Available: https://academic.udayton.edu/kissock/http/Weather/
[29] Tutorials Point. (Accessed: Apr. 2, 2021). *Hadoop—MapReduce*. Tutorialspoint.com. [Online]. Available: https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm
[30] S. Chaudhary. (Accessed: Mar. 20, 2021). *Why 1.5' in IQR Method of Outlier Detection?* Towardsdatascience.com. [Online]. Available: https://towardsdatascience.com/why-1-5-in-iqr-method-of-outlier-detection-5d07fdc82097
[31] K. Muralidhar. (Accessed: Mar. 20, 2021). *Outlier Detection Methods in Machine Learning*. Towardsdatascience.com. [Online]. Available: https://towardsdatascience.com/tagged/z-score?p=1c8b7cca6cb8
[32] Analytics Vidhya. (Accessed: Mar. 25, 2021). *Interpolation—Power of Interpolation in Python to Fill Missing Values*. Analyticsvidhya.com. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/power-of-interpolation-in-python-to-fill-missing-values/
[33] Sas. (Accessed: Mar. 30, 2021). *Predictive Analytics: What it is and Why it Matters*. Sas.com. [Online]. Available: https://www.sas.com/en_in/insights/analytics/predictive-analytics.html
[34] D. Waga and K. Rabah, "Environmental conditions' big data management and cloud computing analytics for sustainable agriculture," *Prime Res. Educ.*, vol. 3, no. 8, pp. 605–614, Nov. 2013.
[35] P. S. Dutta and H. Tahbilder, "Prediction of rainfall using data mining technique over Assam," *Indian J. Comput. Sci. Eng.*, vol. 5, no. 2, pp. 85–90, Apr./May 2014.
[36] Agricoop. (Accessed: Apr. 10, 2021). *Union Territory: Andaman & Nicobar Islands Agriculture. Contingency Plan for District: South Andaman*. Agricoop.nic.in. [Online]. Available: https://www.agricoop.nic.in/sites/default/files/A_N%202-South%20Andaman-03.05.2016.pdf
[37] Krishi. (Accessed: Apr. 15, 2021). *Comprehensive District Agricultural Plan Of Ahmednagar*. Krishi.maharashtra.gov.in. [Online]. Available: http://krishi.maharashtra.gov.in/Site/Upload/Pdf/Nagar_cdap.pdf
[38] WeatherOnline. (Accessed: Apr. 7, 2021). *Port Blair Weather*. WeatherOnline.in. [Online]. Available: https://www.weatheronline.in/India/PortBlair.htm
[39] Climate-Data. (Accessed: Apr. 7, 2021). *Climate: Andaman and Nicobar Islands*. Climate-Data.org. [Online]. Available: https://en.climate-data.org/asia/india/andaman-and-nicobar-islands-2291/
[40] D. S. Zingade, O. Buchade, N. Mehta, S. Ghodekar, and C. Mehta, "Crop prediction system using machine learning," *Int. J. Advance Eng. Res. Develop., Recent Trends Data Eng. Crop Predict. Syst. Mach. Learn. All Rights Reserved Sci. J. Impact Factor*, vol. 4, no. 5, pp. 1–6, 2017.
[41] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," in *Proc. 13th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2016, pp. 1–5.
[42] W. A. Goya, M. R. de Andrade, A. C. Zucchi, N. M. Gonzalez, R. de Fatima Pereira, K. Langona, T. C. M. de Brito Carvalho, J.-E. Mangs, and A. Sefidcon, "The use of distributed processing and cloud computing in agricultural decision-making support systems," in *Proc. IEEE 7th Int. Conf. Cloud Comput.*, Jun. 2014, pp. 721–728.
[43] S. Brdar, J. Crnobarac, D. Culibrk, B. Marinković, and V. Crnojević, "Support vector machines with features contribution analysis for agricultural yield prediction," in *Proc. 2nd Int. Workshop Sensing Technol. Agricult., Forestry Environ. (EcoSense)*, Apr. 2011, pp. 43–47.
[44] N. Gandhi, O. Petkar, and L. J. Armstrong, "Rice crop yield prediction using artificial neural networks," in *Proc. IEEE Technol. Innov. Agricult. Rural Develop. (TIAR)*, Jul. 2016, pp. 105–110.
[45] A. K. Sharma, G. Aggarwal, S. Bhardwaj, P. Chakrabarti, T. Chakrabarti, J. H. Abawajy, S. Bhattacharyya, R. Mishra, A. Das, and H. Mahdin, "Classification of Indian classical music with time-series matching deep learning approach," *IEEE Access*, vol. 9, pp. 102041–102052, 2021.
[46] S. Tiwari, A. Jain, A. K. Sharma, and K. Mohamad Almustafa, "Phonocardiogram signal based multi-class cardiac diagnostic decision support system," *IEEE Access*, vol. 9, pp. 110710–110722, 2021.

RISHI GUPTA (Member, IEEE) received the M.Tech. degree in software engineering from Rajasthan Technical University, Kota, Rajasthan, in 2013, and the Ph.D. degree in computer science and engineering in the area of face recognition in image processing from Jagannath University, Jaipur, Rajasthan, in 2019. He has ten years of teaching experience. He is currently an Assistant Professor with the Department of Computer Science and Engineering & IT, Manipal University Jaipur. His current research interests include image processing and machine learning.

AKHILESH KUMAR SHARMA (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in computer science and engineering. He is currently working with Manipal University Jaipur, Rajasthan, India, as an Associate Professor. He has more than 18 years of experience. He has been chairing sessions and as an Expert for keynotes in IITs, NITs, Vietnam, Thailand, Malaysia, Australia, China, and Singapore. He has presented many research articles in international journals and conferences and organized various FDP's, events, conferences, and workshops. He holds four patents and four copyrights to his credit and has setup cognitive intelligence research lab in Jaipur. His research interests include the area of soft computing, machine learning, bigdata analytics, and healthcare. He is affiliated with IEEE, ACM, CSI, (IUCEE), and MIR Lab, USA. He is currently the Joint Secretary of ACM Professional Chapter Jaipur.

OORJA GARG was born in Lucknow, Uttar Pradesh, India, in 2000. She is currently pursuing the B.Tech. degree in computer science and engineering with the School of Computing and Information Technology, Manipal University Jaipur, Jaipur, Rajasthan, India. She is also interning as a Developer at Glorich India Pvt., Ltd. Her research interests lie in the field of big data analytics and utilizing it to develop and improve predictions concerning the agricultural sector.

KRISHNA MODI was born in Mumbai, Maharashtra, India, in 2000. She is currently pursuing the bachelor's degree in technology in computer science and engineering with the School of Computing and Information Technology, Manipal University Jaipur, Jaipur, Rajasthan, India. She is currently interning as a Developer at a financial service company VenEx, India. Her research interests include the development of prediction and recommendation systems in agricultural sector using big data analysis, fundamental concepts of weather prediction systems using meteorological data, and understanding development of smart farming concepts.

SHAHREEN KASIM is currently an Associate Professor with the Department of Security Information and Web Technology, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. Her areas of interests include bioinformatics, soft computing, data mining, and web and mobile applications.

ZIRAWANI BAHARUM received the B.Sc. degree in computer science majoring in modeling and industrial computing, the M.Sc. degree in information technology, and the Ph.D. degree in computer science from Universiti Teknologi Malaysia (UTM), in 2003, 2005, and 2017, respectively. She is currently a Senior Lecturer with the Technical Foundation Section, Universiti Kuala Lumpur, Malaysian Institute of Industrial Technology (UniKL MITEC). Her research interests include computer modeling and simulation, integrated model development, computer science, and information and communication technology (ICT).

HAIRULNIZAM MAHDIN (Member, IEEE) received the Ph.D. degree from Deakin University, Australia, in 2012, and also completed his Ph.D. thesis in the same year. He is currently an Associate Professor with the Department of Information Security and Web, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. His research focuses on the area of data management, the IoT, RFID, information security, software engineering, and web technology. He is currently working as the Deputy Dean of Research, Development, and Publication.

SALAMA A. MOSTAFA received the B.Sc. degree in computer science from the University of Mosul, Iraq, in 2003, and the M.Sc. and Ph.D. degrees in information and communication technology from Universiti Tenaga Nasional (UNITEN), Malaysia, in 2011 and 2016, respectively. He is currently a Lecturer with the Department of Software Engineering, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). His research interests include the area of soft computing, data mining, software agents, and intelligent autonomous systems.