EasyChair Preprint

№ 7443

WCP: Weather-Based Crop Yield Prediction Using Machine Learning and Big Data Analytics

Sridev Bhavanandam

February 8, 2022

WCP: WEATHER-BASED CROP YIELD PREDICTION USING MACHINE LEARNING AND BIG DATA ANALYTICS

Sridev Bhavanandam

Abstract

Agriculture is the Indian economy's backbone. Big data analytics are becoming more precise and feasible in agricultural research. Current water scarcity, uncontrollable costs owing to demand-supply imbalances, and weather instability need farmers to be prepared with smart farming techniques. Crop yields must be addressed due to unknown climate changes, limited irrigation infrastructure, soil fertility decrease, and conventional agricultural approaches. In agriculture, machine learning (ML) is used to forecast crop output. Many ML approaches such as prediction, classification, regression, and clustering anticipate agricultural production. We presented the WCP approach for predicting agricultural yields based on climatic variables in Big data analytics. The proposed study proposes a crop recommendation system that employs MapReduce and improved K-means (IKM) clustering to get computationally efficient results. The MapReduce framework may be used to build MapReduce and crop prediction based on meteorological conditions by employing Categorization, Attribute Selection, C5.0, and association algorithms. Choosing the proper method from the pool of available algorithms, on the other hand, offers a problem to the researchers in terms of the crop. This work looks at how different machine learning techniques may help estimate agricultural production. A method for predicting agricultural production using ML algorithms in the big data computing paradigm has been suggested. This report also includes a study of ML algorithms for large data analytics.

Keywords: Bigdata, Weather, Crop, Prediction, Machine Learning, MapReduce, K-Means

I. INTRODUCTION

Farmers and agriculture throughout the nation suffer from abrupt weather conditions as they fail to produce adequate crops. India is an agricultural nation with the world's second-largest cultivable land area of more than 1.6 million square kilometers. Several key businesses in India rely on agriculture for raw materials, including the cotton and jute textile industries, sugar, Vanaspati, and others [1]. There is no such thing as a universal agricultural assistance scheme [2]. India is a developing nation that is heavily reliant on agriculture. Despite having a large amount of digital data, they cannot obtain real-time factual information such as crop yield statistics, soil and crop disease detection methods, pesticides to be applied, weather conditions, pest management, etc. As a solution to increase usability tools, this article investigates constructing a solution that seeks to be scalable, simple to access, community-oriented design, and efficient to bridge the digital divide between rural farmers and technology [3][4].

The weather has a significant impact on agricultural productivity. It has a significant impact on crop growth, development, yields, the occurrence of pests and diseases, water requirements,

and fertilizer requirements [5]. This is due to variances in nutrient mobilization due to water stress and the timing and efficacy of preventative measures and crop cultural activities. Weather anomalies may cause physical harm to crops and soil erosion [6]. The quality of agricultural production relies on the weather as it moves from the farm to storage and then to market. Bad weather may impact product quality during transportation and the viability and vigor of seeds and planting material during storage. As a result, no part of crop cultivation is immune to weather effects. Weather conditions influence crop growth, development, and yield [7].

They also contribute to the occurrence and spread of pests and illnesses. The susceptibility of crops to weather-induced stressors and pests and diseases varies between crops, varieties within the same crop, and growth phases within the same crop variety [8]. Weather factors demonstrate geographical changes in a region at a particular time, temporal fluctuations at a given location, and year-to-year variations for a given place and time, even on a climatological basis [9]. The weather during short intervals and year-to-year changes at a specific location across the given time frame must be considered for cropping purposes. The coefficient of variability, defined as the percentage deviations of extreme values from a mean or median value for any given time unit, measures the parameter's variability [10].

The higher the degree of fluctuation of a specific weather characteristic, the shorter the time unit [11]. The severity of the three variants mentioned above varies depending on the Range of meteorological parameters. Rainfall is the most changeable of all time and space characteristics during short periods [12][13]. The short-period interannual variability in rainfall is high, which implies that variability must be represented in terms of the % likelihood of receiving a particular quantity of rain. The minimum guaranteed rainfall levels must be indicated at a given probability level. Crops and cropping techniques must be designed. Although their cardinal phased weather needs meet the temporal march of the appropriate weather element(s), endemic times of pests, illnesses, and hazardous weather are avoided. Short-period meteorological data, both routine and processed (such as initial and conditional probability), play an important part in such strategic crop and cropping practice planning [14].

This study creates a crop recommendation system using MapReduce and IKM clustering, which yields efficient computing results. The model focuses on a broad variety of crops and their yield per area, soil type, and seed types based on the kinds utilized in a given location.

The remaining paper is organized. Section 2 discusses related work about Crop yield prediction, Section 3 discusses the WCP methodology and Bigdata MapReduce, Section 4 discusses the results and discussion, and Section 5 discusses the conclusions.

II. BACKGROUND STUDY

P. S. Cornish et al. [2] described the architecture of the open platform for big data analytics used to improve crop yield for food security and better lifestyles across a variety of crop alternatives and cropping systems. Farmers plant rain-fed crops in the Kharif to accommodate pre-Kharif cultivation and the Rabi winter. They had access to some irrigation. How to Reduce Climate Risk in a High-Risk

Environment using Rain-fed Rice Farmers' systems is critical for water resources; consequently, we advocate replacing rice fallow with a more secure climate-responsive strategy for watershed development. Rabi vegetable crops may need the use of tiny water collection systems. The approach evaluated less hazardous agricultural practices that produced more water.

By lowering the dimensionality of data, K. Sabarina and N. Priya [3] devised an efficient approach for Bigdata for precision agriculture. By collecting and evaluating real-time data, predictive analytics can assist farmers in making the best choices possible. With the exponential growth in the volume of big data, the employment of a tensor-based feature radio model posed a challenge to data processing efficiency. Big data is one way to enhance data analysis performance on weather, soil, air quality, crop maturity, and labor costs. It is especially critical in precision agriculture, where it handles real-time data processing and streaming data. Real-time data were gathered from a variety of agricultural sources.

MR Bendre et al. [4] advocate for web-based information in agriculture and weather forecasting for future farming. The employment of a programming model and a complexly distributed algorithm for data processing results in applying big data analytics to future processes and challenges in agricultural prediction. Agriculture applications have a unique opportunity to provide advanced weather to increase crop output and decrease unnecessary harvesting costs. Precision agriculture's future uses and challenges stemmed from using a programming model and distributed algorithm for data processing and weather forecasting. The model illustrates the temperature and precipitation in the region as a result of this conclusion. Crop patterns and irrigation management are two strategies for boosting output and profit.

J.W. Kruize and colleagues [5] developed a farm software ecosystem for smart farming, an architecture for monitoring, planning, and controlling agricultural processes, and a smart farm reference architecture for analyzing the design and implementation of farm software ecosystems. Configure the different components more effectively. Ecosystems of agricultural software contribute to the development of software that allows intelligent farming.

Ahrary and D. Ludena [6] presented a profile-based precision agriculture architecture to facilitate service-based real-time decision-making. The Internet of Things is being used to automate the process of developing integrated environment information research in agricultural and commercial sectors to provide a richer user experience and relevant information about the unique environment. Numerous benefits of bigdata healthy food recommendation to the system's end-user and different analytics to improve system performance are critical for the nutrition-based vegetable production and distribution system. Farmers may profit from cost-effective and simple-to-implement technology solutions.

S. S. Reddy and C. S. Bindu [7] for developing a strategy for agricultural integrated analytics. Advances in internet speed have enabled data to be transported at incredibly rapid speeds on a global scale. A vast amount of data is continually being generated in the form of data streams from various real-world

applications. Clustering algorithms for big data analysis Clustering is an unsupervised learning approach used to classify massive data sets with similar features. Clustering data streams poses extra challenges, such as managing limited time, memory, and noisy data, as well as clustering high-dimensional data.

S. Athmaja et al. [8] presented effective solutions for huge data analytics in studying machine learning algorithms. Agricultural communities worldwide have benefited from comparison information produced from big data analysis; via machine learning algorithms, agricultural communities have gained some comparative knowledge and adjustments to traditional agriculture.

P. Shah et al. [10] demonstrated a big data analytics architecture for an agricultural advisory system and constructed an analytical engine using open-source frameworks. Agro-advice systems that use big data analytics can increase agricultural productivity. The research is used to reduce technical communities and information through ideas and decision support systems to propose and create an open-source, cost-effective, and scalable big data analytics architecture for an agreed-upon system agricultural production prediction. Farmers should project yields based on current weather conditions and decide whether or not to produce that particular crop as a substitute crop if the forecasted yield is negative.

R. Kaur and colleagues [11] stressed the need to build a framework for detecting crop disease based on historical symptoms and recommendations. Big data analysis frameworks need machine learning approaches because massive messages cannot be stored in traditional data structures. The Hadoop platform, written in Java, collects this data and establishes a framework for sickness recommendation solutions. The big data analytics technology detects crop disease and recommends a solution based on symptom similarities. Because conventional techniques are incapable of storing and interpreting this amount of data, a data-parallel processing, and analytical paradigm are required. Agriculture is establishing a framework. To do this, identify illnesses based on their symptoms and associated treatments based on their high degree of similarity. The Hadoop and Hive technologies were used to collect, clean, and standardize the data.

S. Rajeswari et al. [14] offered a conceptual design of a big data open platform to assist various industries, including agriculture. Big data, mining techniques, cloud-based big data analytics, and IoT technologies play major roles in smart agricultural feasibility studies. Precision agricultural systems are expected to play a critical role in agricultural activity improvement.

III. WCP METHODOLOGIES

The major objective would be to analyze the data using MapReduce and write a Python recommender algorithm to extract output depending on seasonal conditions and geography, then perform IKM clustering and determine the mean production per area provided by a group of crops in a certain location. We picked our system's temperature, rainfall, wind speed, humidity, soil type, and seed type as decision factors based on past work in other journals. Python will be used to collect and preprocess the raw data. The preprocessed data is then utilized for Hadoop's MapReduce framework to

process the data. MapReduce is a programming technique that uses a parallel, distributed method to handle massive volumes of data.

Following that, we suggest combining all of the MapReduce datasets for the various parameters into a single final/super dataset and developing a recommendation system. The month, area, and state may all be entered by the user. After that, we'll establish numerous agricultural seasons. After that, we assign the month supplied by the user to the proper season/seasons. For instance, if November is specified as the month, we may assign Rabi/Winter crops. We then analyze the data to determine the three crops that yield the most during that season and the crops that yield the most throughout the year in that state and location. These would be stored in two separate data frames, one for each season and another for the whole year.

Additionally, we output the temperature, rainfall, wind speed, and humidity values previously given by the crop. Additionally, we list the seed variety for each soil type for crops grown in different areas since seed availability varies by region. The recommendation function's output will be shown on a graphical user interface (website) created using Flask and Python. The user may submit pertinent data and get a response from the system.

Following that, we will use an IKM clustering technique. First, we will construct an elbow graph to establish the cluster count and the required K value. This will be accomplished via the usage of the Scikit-learn package. Following that, the fit predicts approach will be utilized to get cluster values. This will be accomplished via an array, where numbers starting with 0 indicate the values of a single cluster. The clusters will then be plotted using the scatter method provided by the Matplotlib toolkit. Each cluster centroid will be exhibited to reflect the average value of the cluster in which each crop is plotted, and a separate color will represent each cluster.

3.1 DATA COLLECTION:

The production of crops. All of the states and their districts were included in the CSV file. It offers information on 125 crops, their productivity, and the region in which they were planted from 2000 to 2014 for six seasons: Kharif, Rabi, Summer, Winter, Autumn, and the whole year. The daily average temperature of the cities from 1995 to 2020 is included—India's rainfall from 1901 to 2015. From 1901 to 2015, CSV contains the average monthly rainfall of several subdivisions. The soil, seed, humidity, and wind speed data were manually gathered from several sources such as agricoop, the Department of Agriculture's weather online, and climate-org.

3.2 Preprocessing

This stage included merging and cleaning the collected datasets. We uploaded our datasets to the Colab notebook and utilized pandas data frames to eliminate superfluous columns while retaining the important ones. We used the NumPy and SciPy libraries to do our calculations. A few index columns were added for future calculations. Interpolation, as shown, was used statistically to get the estimated value for the dataset's missing value.

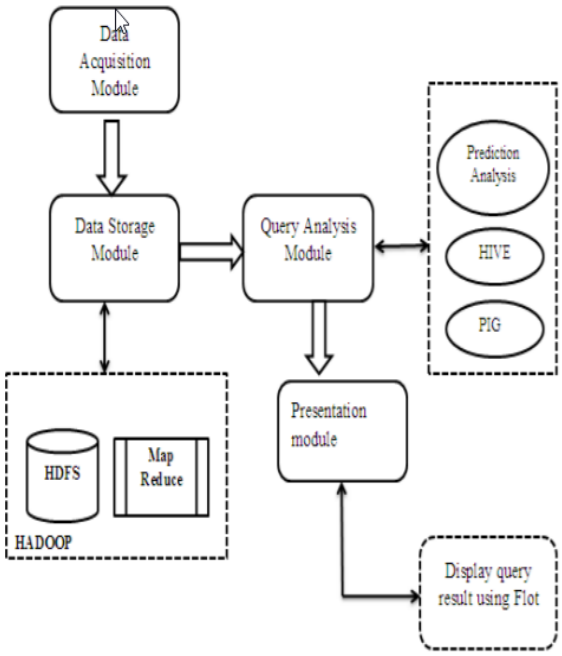

Figure 3.1 System Architecture

3.3 Data Acquisition Module

This module collects sensor data, weather forecasts, social media, and market trends. These meteorological data can be issued manually or acquired using meteorological data acquisition equipment; small received data are initially stored in an Oracle database; once a sufficient number of small data has been gathered, small data are transferred to the storage module; transferred data are automatically deleted.

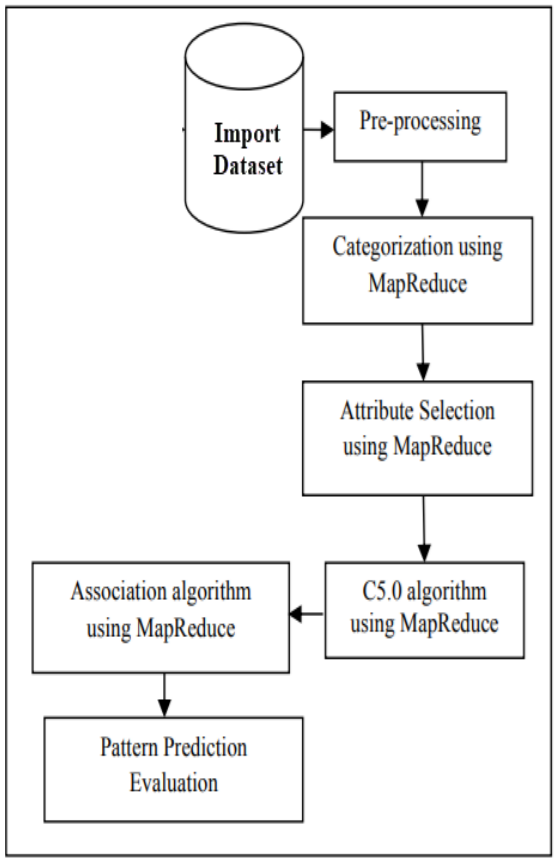

Figure 3.2 Level -0 Data Flow Diagram

3.4 Data storage module: It is responsible for storing metadata and data sets in duplicate, acting as a backup facility. HDFS is a data-agnostic storage container. Small data collected in the data collection module will be saved in the storage module regularly after accumulating to a certain amount.

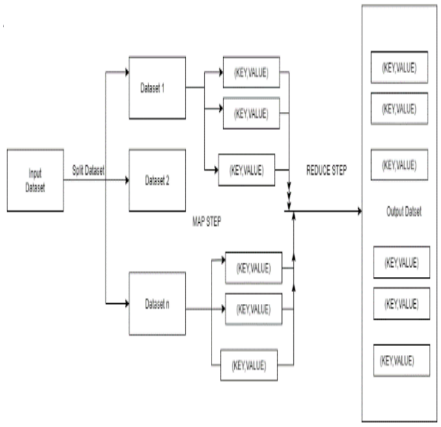

Figure 3.3 Level -1 Data Flow Diagram

3.5 Query Analysis: This module's processing phase consists of two parts: data reading/analysis and forecast outcomes. Hive is primarily responsible for data reading. Hive is a Hadoop-based data warehouse architecture. It was designed to allow analysts with good SQL expertise to operate on the massive amounts of data contained in HDFS. Hive is a workstation application

that turns SQL queries into a sequence of MapReduce tasks executing a Hadoop cluster. MapReduce is an execution engine designed for big data processing that greatly reduces query response time. The second section includes a prediction function that uses the IKM cluster technique to generate forecast data. We utilize apache mahout in this case. It is an open-source, scalable machine learning library. Mahout is a fast technique to create unsupervised machine learning algorithms. The data from the last several years are utilized to forecast the future.

3.6 Presentation: This module will show the findings produced from the query analysis module. The ability to represent complicated data using charts and graphs is critical for data analysis. Figure 3 depicts the suggested system process in the first phase, whether datasets acquired from different sources are further preprocessed to enter the prediction algorithm effectively. After cleaning the data, put it into HDFS and run a Hive query to analyze it. We can also execute a Py script to analyze the data, and the output is sent to flotend to generate the graph for examination. We utilize the logistic regression approach for prediction, which can be implemented using the Apache Mahout ML package.

3.7 An algorithm in Hadoop Theme:

We proposed a technique in the study that assists us in predicting crop production by recommending the optimal crop. It also focuses on the soil type, which helps predict which crop has to be planted in the field to increase productivity.

Model: Soil types are important in terms of agricultural productivity. The soil information may be acquired by considering the previous year's weather. It assists us in predicting which crops are suited for certain climatic circumstances. Crop quality may also be improved by using weather and disease data sets. A substantial value of the data set may be used to forecast the crop. An Agro algorithm can handle a vast number of data sets. It is implementable on the Hadoop platform and uses the Hadoop framework to manage massive data sets.

3.8 Observation:

According to the research discussed above, varied soil samples obtained from various sites may be assessed utilizing portable NPK sensors with a short response time. The NPK sensor, on the other hand, will detect just nitrogen, potassium, and phosphate in the soil. Numerous authors' work has been combined in one area to aid researchers in comprehending the current state of agriculture. The strategy allows us to make comparisons between agricultural approaches and crop compositions. The Agro algorithm supports farmers in making crop selection decisions. Because climatic conditions are subject to change, trustworthy conclusions are impossible. It assists the farmer in planting the proper crop accurately and quickly. The forecast is based on the atmosphere, incorrect since climatic circumstances might change.

3.9 Improved K-Means Clustering for Crop Yield Prediction

Improved K- Means clustering is one of the most effective methods for predicting crop productivity. The key benefits of employing this approach are that it may have numerous parameters and, if the K value is modest, it may be computationally closer to hierarchical clustering. IKM may

create more stubborn clusters than hierarchical clustering, especially when the clusters are rotund. IKM characteristics are always K clusters. Each cluster contains at least one item. The clusters are non-hierarchical and expand beyond the boundaries. Because nearness does not entail the 'center' of clusters, each cluster member is closer to its cluster than another. The IKM cluster technique partitions the dataset into K clusters. The data points are randomly assigned to the clusters, resulting in nearly the same data points. Put it down if the point is closer to its cluster for each data point. If the data point is close to a cluster, it should be moved to the closest cluster. Repeat the procedures until the sum of the data points surpasses the sum, resulting in no point influencing one cluster over another. At this point, the clusters are consistent, and the processes have halted.

3.10 Improved K-means Clustering Algorithm

The IKM clustering algorithm is a partitioning technique that varies in determining the number of clusters and choosing the first cluster centroid. The following are IKM algorithms' processing of multiple clusters and the initial cluster centroid.

The IKM technique is designed to identify initial cluster centers and many clusters. The approach begins by selecting the attribute with the least value in the X dataset. After deducting the minimum value from each record and adding it to the removed values, a single representative value for each record is generated. Sorting the whole collection of representative values is required for partitioning. A series of iterations accomplish the partitioning. Each partitioned data value is retained for further computation of partitioned centroids. The IKM function is passed the X dataset, the iteration number, and the value of the centroids. The steps of the IKM function are as follows.

1. The Euclidian distance is computed of each point from X with centroids value, and the minimum distance value is assigned to the cluster.
2. The center value is moved to the position by computing the mean of all cluster points. This is repeated until the centroid value gets the same.
3. The output value is acquired in tens of cluster index and centroid value. For cluster representation, the silhouette function is used for plotting those points.

The mean value of the clustered index is then calculated to get a compact value. This is repeated till the number of iterations has been reached. Finally, the highest mean of all means is calculated, and its index value is established as several clusters. The sorted representative value is partitioned again based on the index value, and the same method is repeated up to the IKM function to provide refined output values.

IV. RESULTS AND DISCUSSION

The WCP method has been implemented using Python programming and Bigdata Hadoop Environment. The WCP method achieves the highest accuracy than existing ones.

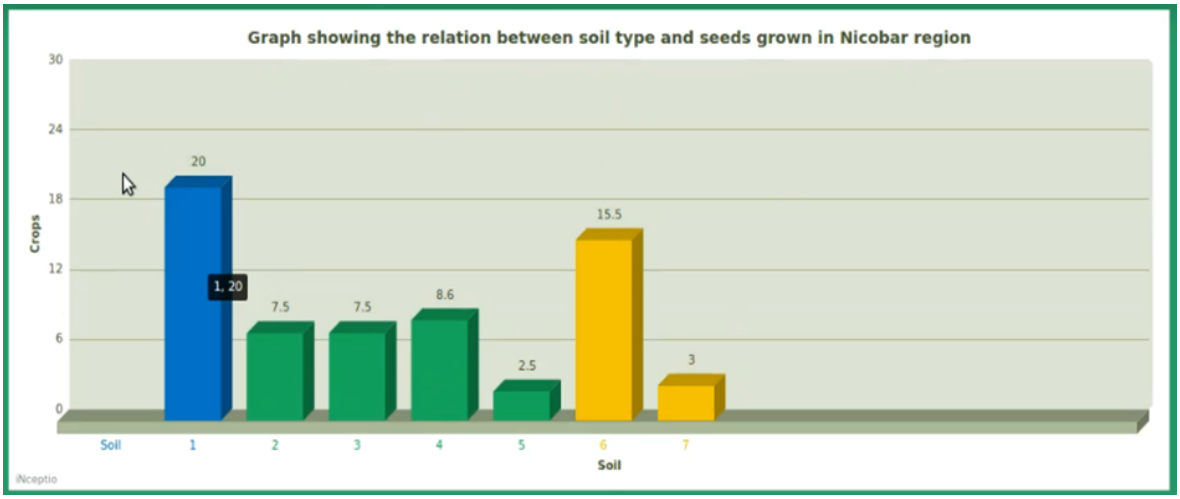

Figure 4: Relation between soil types and seeds

Figure 4 represents the relation between soil type and seeds grown in the Nicobar region. In x-axis denotes the soil, and the y-axis denotes the crops.

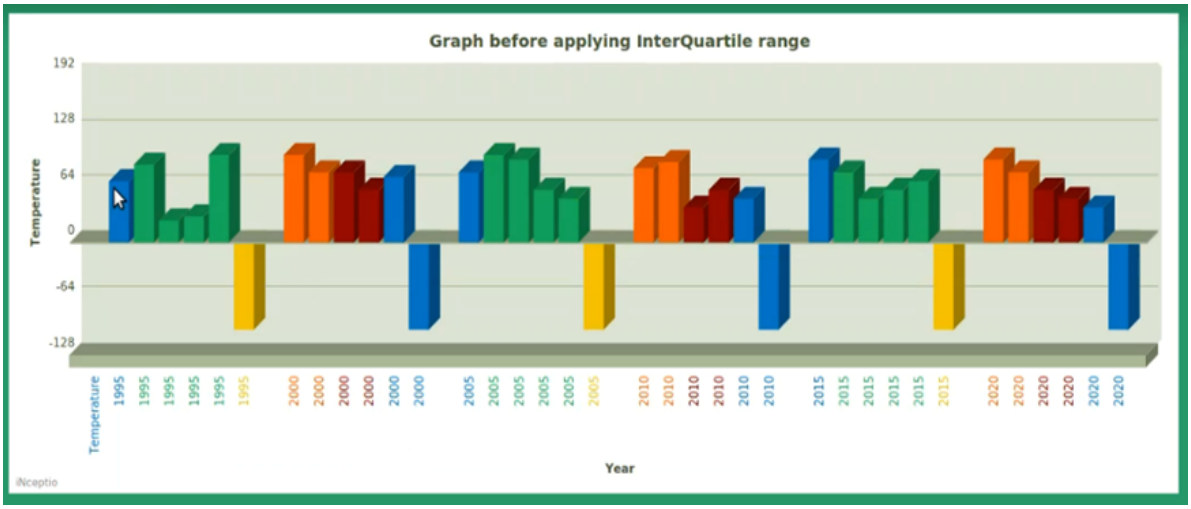

Figure 5: Before applying the inter-quartile Range

Figure 5 represents the dataset values before applying the inter-quartile range data's

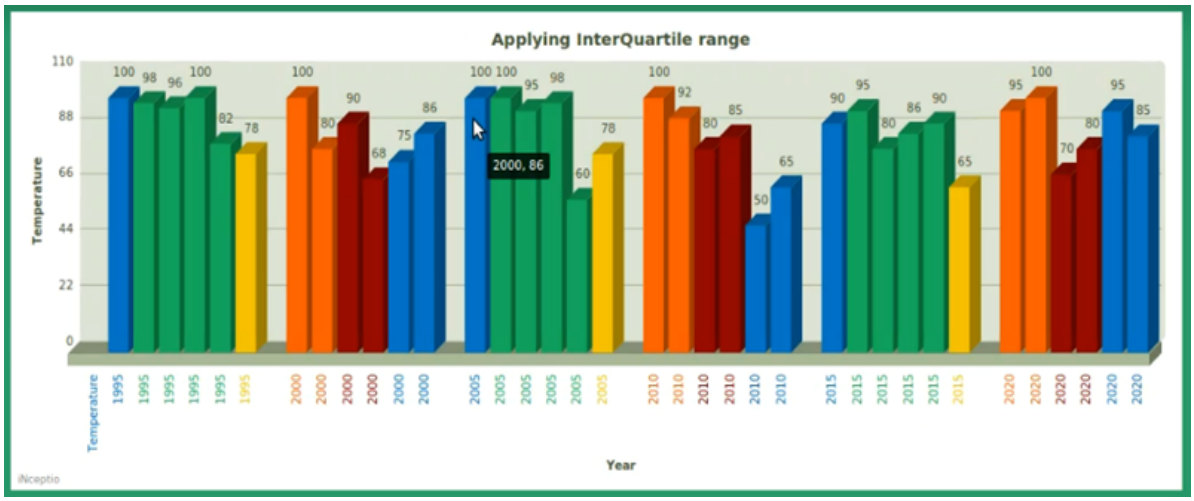

Figure 6: Applying inter-quartile Range

Figure 6 represents the various years and the temperature levels between 1995 to 2020

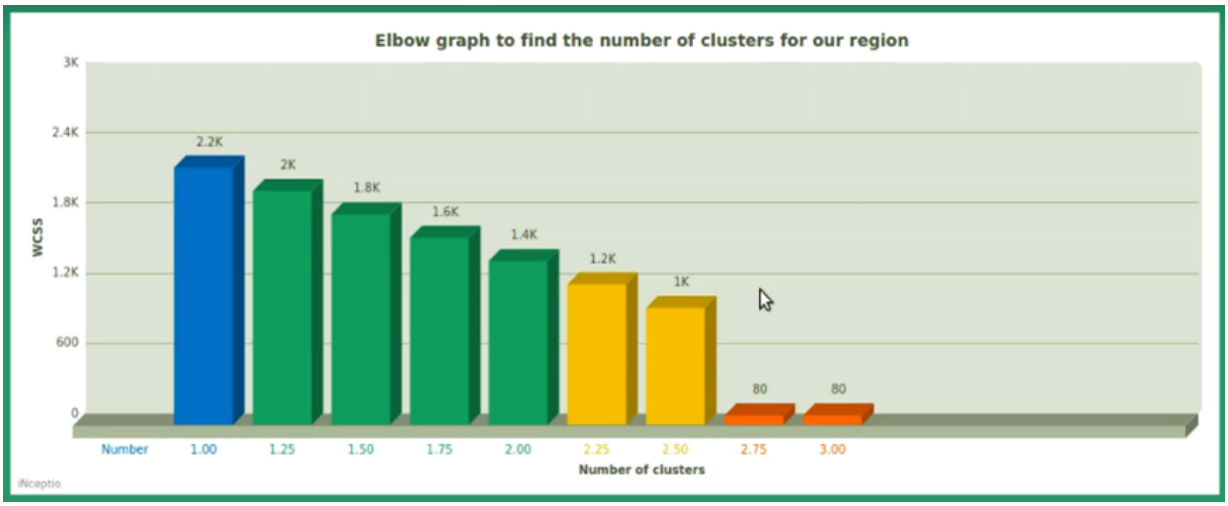

Figure 7: clusters in our region

Figure 7 represents the weather-based crop yield data's to cluster in various levels using the IKM clustering algorithm.

V. CONCLUSION

We demonstrated the WCP technique, incorporating a crop recommendation system, MapReduce, and IKM clustering to provide computationally efficient results. The model considers various crops and their products according to location, soil type, and seed types depending on the varieties grown in a particular area. The mean production of a set of crops may be calculated using IKM clustering visualization graphs. The recommender function and IKM Clustering algorithms are accessible on https://github.com/oorjagarg/WB-CPI. The link between elements (such as the optimal temperature, seasonal rainfall, wind speed, humidity, soil availability, and seed types required), crop, and the area has also been researched and shown using 2D and 3D graphs. The strategy is scalable, and in the same manner that the methodology is discussed, it may be used to determine the proposed crops for various states.

VI. REFERENCES

[1] Gao, "The impact of climate change on China's crop production: A CMIP5 ensemble assessment," 1st Int. Conf. Agro-Geoinformatics, Agro-Geoinformatics 2012, pp. 208–212, 2012.

[2] P.S. Cornish et al., "Improving crop production for food security and improved livelihoods on the East India Plateau II. Crop options, alternative cropping systems, and capacity building," Agric. Syst., vol. 137, pp. 180–190, 2015.

[3] K. Sabarina and N. Priya, "Lowering data dimensionality in big data for the benefit of precision agriculture," Procedia Comput. Sci., vol. 48, no. C, pp. 548–554, 2015.

[4] MR Bendre, R.C. Thool, and V.R. Thool, "Big data in precision agriculture: Weather forecasting for future farming," Proc. 2015 1st Int. Conf. Next Gener. Comput. Technol. NGCT 2015, pp. 744–750, September 2016.

[5] J.W. Kruize, J. Wolfert, H. Scholten, C.N. Verdouw, A. Kassahun, and A.J.M. Beulens, “A reference architecture for Farm Software Ecosystems,” Comput. Electron. Agric., vol. 125, pp. 12–28, 2016.

[6] Ahrary and D. Ludena, "Research Studies on the Agricultural and Commercial Field," Proc. - 2015 IIAI 4th

Int. Congr. Adv. Appl. Informatics, IIAI-AAI 2015, pp. 669–673, 2016.
[7] KSS. Reddy and CS Bindu, "A review on density-based clustering algorithms for big data analysis," 2017 Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), pp. 123–130, 2017.
[8] S. Athmaja, M. Hanumanthappa, and V. Kavitha, "A survey of machine learning algorithms for big data analytics," 2017 Int. Conf. Innov. Information, Embed. Commun. Syst., pp. 1-4, 2017.
[9] J. Majumdar, S. Naraseeyappa, and S. Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," J. Big Data, vol. 4, no. 1, 2017.
[10] P. Shah, D. Hiremath, and S. Chaudhary, "Big data analytics architecture for an agro advisory system," Proc. - 23rd IEEE Int. Conf. High Perform. Comput. Work. HiPCW 2016, pp. 43–49, 2017.
[11] R. Kaur, R. Garg, and H. Aggarwal, "Big data analytics framework to identify crop disease and recommendation a solution," Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016, vol. 2, 2017.
[12] H. Jain and R. Jain, "Big data in weather forecasting: Applications and challenges," 2017 Int. Conf. Big Data Anal. Comput. Intell. pp. 138–142, 2017.
[13] P. Shah, "Towards Development of Spark Based Agricultural Information System including Geo-Spatial Data," pp. 3394–3399, 2017.
[14] S. Rajeswari, K. Suthendran, and K. Rajakumar, "A smart agricultural model by integrating IoT, mobile and cloud-based big data analytics," Proc. 2017 Int. Conf. Intell. Comput. Control. I2C2 2017, pp. 1–5, Jan. 2018.