2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) | 978-1-6654-7467-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSSIT55814.2023.10060865

A Study on Weather based Crop Prediction System using Big Data Analytics and Machine Learning

Santhoshini Sahu
Department of CSE
GMR Institute of Technology
Rajam, Andhra Pradesh, India
santhoshini.s@gmrit.edu.in

T. Daniya
Department of Information Technology
GMR Institute of Technology
Rajam, Andhra Pradesh, India
daniya.t@gmrit.edu.in

R. Cristin
Department of CSE
GMR Institute of Technology
Rajam, Andhra Pradesh, India
cristin.r@gmrit.edu.in

***Abstract* - Agriculture plays a key role in the Indian economy. There are various challenges in the agricultural sector posed by climatic conditions, loss of biodiversity, soil erosion, traditional farming working with plants, diseases, and pests. The impact of using conventional methods of agriculture on the environment include deforestation, soil erosion and depletion of nutrients in the soil. Indian farmers without having any prior knowledge about the atmospheric conditions cannot choose their crops effectively. Incorporating modern agriculture can provide many solutions to these kinds of problems. Precision farming techniques can solve this issue by gathering and evaluating data on temperature, rainfall, soil, seed, crop production, humidity, and wind speed, which will assist farmers in raising agriculture production. The data is refined before being examined and processed using the map reduce architecture. Applying the K-means clustering method to the map reduce results yields a mean accuracy result for the data. Taking Consideration of regions, the relationship between different factors is analyzed using bar graphs and scatter plots which is hugely beneficial for crop prediction. This study gives an analysis of different crop prediction techniques.**

Keywords: Agriculture, Big Data Analysis, Visual Representation, K-Means Clustering, and Map Reduce.

INTRODUCTION

One of the key areas of public concern is related to agriculture. The production of a crop relies mainly on the quality of the soil, seeds, climatic conditions etc., because it supplies a large amount of food but many nations still go hungry today owing to a shortage or lack of food with an increasing population. Crop production forecasting is a crucial and responsible component in raising food security. Indian farmers frequently struggle with the issue of choosing crops without considering the soil's requirements. By obtaining information and monitoring weather, precision agriculture can solve this issue. Temperature, humidity, soil, seed, crop productivity, and statistics about wind speed that will be useful to farmers increase the yield of crops. Basically, by pre-processing the data the map Reduce analyses and transforms a substantial amount of data. K-means clustering algorithm works on the outcomes of Map Reduce calculates the mean value on the accuracy of the data. The link between different factors is then studied using bar graphs and scatter plots. Studies that are helpful for making predictions of the crop. This study focuses on different crop prediction techniques and how area effects the crop prediction and climatic conditions which effect the crops.

LITERATURE SURVEY:

In the subfields, monitoring of crops at very high resolution severs using spatiotemporal image fusion. Precision crop management can only be effectively implemented with subfield-level or satellite-based time-series crop monitoring. Spatiotemporal image fusion methods that are presently available are useful but however, were frequently suggested to produce medium resolution pictures. Study suggests a high-resolution spatiotemporal image fusion approach (also known as HISTIF) made up of multiplicative modulation of temporal change (MMTC) and filtering for cross-scale spatial matching (FCSM). Heterogeneity, spatial-temporal fusion, and image fusion are the major concepts used in this research. The goal of spatial-temporal fusion is to combine images with different levels of resolution to create an image with sufficient temporal and spatial resolution.

Precision planting, effective management, wise decision-making, and quantitative implementation are becoming increasingly crucial in modern agricultural production.[1]

Crop yield estimation is crucial for quantitative and financial field evaluation to establish strategic goals in agricultural products for export and import programs, which also boosted farmer earnings. This study discusses how to anticipate crop yield using machine learning to increase palm oil production prediction. The primary goals here is to investigate the potential outcomes of forecast of yielding of palm oil based on machine learning, the areas where remote sensing is applied, disease mapping, illness identification, and plant growth as well as tree counting, ideal features, and algorithms have received a lot of attention considering critical assessment of related existing research. A potential architecture palm oil yield prediction has been developed using machine learning. Studies analyzing agricultural yield forecasting and the creation of a highly effective model for the most accurate forecast of these yields minimally challenging in terms of computation. Frameworks for learning provide a clear understanding of the method of evaluating the enormous amounts of data and analyzing the data that was gathered. The models outlining the links between. Through this, actions and constituents are constructed technologies. Consequently, a linear SVM, a regularized classifier has shown better performance than LDA which is a non-regularized one. The process of training of large datasets is complicated and time-consuming when traditional methods of machine learning are used. [2]

One prospective area for research is the prediction of crop output based on environmental, soil, water, and crop characteristics. Models based on deep learning are very frequently used in the extraction of important crop characteristics for forecasting. Though these techniques might address the yield forecast. Several deficiencies which include the following no direct non-linear or linear model could be built mapping between crop yield and the raw data values, as well as how those models perform heavily depends on the extracted material's quality features. The benefits of deep reinforcement learning giving the things direction and motivation shortcomings. Bringing together the knowledge of deep learning along with reinforcement learning, deep using reinforcement learning, a full crop is produced. A system for yield prediction was designed that can map the raw data values for crop forecast.[3]

Utilizing hierarchical components for crop yield prediction using multiherbal Gaussian and 3D Convolutional Neural Networks Process. Numerous crop simulations are available. Although there exist models, they only apply to just in a particular location. To get around this limitation based on some methods of machine learning later construct a model based on the characteristic gleaned from images from distant sensing. Timely and exact remote sensing-based crop yield forecasting Data are crucial for ensuring food security. Though, Crop yield is a difficult process, making. It is really a challenging task to do better. A novel 3-D convolutional multichannel network is suggested to capture characteristics in an order for this challenge using agricultural yield forecasting. A complete 3-D convolutional a neural network is built to maximize investigation of deep spatial-spectral featured spectral pictures. Afterward, a multichannel a proposed learning (MKL) strategy for combination of deep spatial-spectral features inside a picture, along with the aspects of spatial consistency between samples. We designate a set of nonlinear each feature of the MKL framework has its own kernel, it offers a reliable method to fit features taken from several domains.[4]

Use machine learning to anticipate crops based on environmental parameters. This study uses the algorithms Random Forest and XGDBoost algorithms. Farming heavily depends on Rainfall is dependent on a variety of soils, parameters, specifically nitrogen and phosphorus potassium with environmental conditions like temperature both rain and the development of technology in crop productivity will rise as a result of agriculture. Smart farming is using remote sensor embedded systems like IoT devices and lot more.

Machine learning is an exciting area of study for crop forecasting relied on data trends. The suggested system will be incorporated with sensors such as PH sensors, moisture sensors, rain sensors, sensors for monitoring the humidity and temperature data gathered from such sensors and machine algorithms for learning: Random Forest and Boost. The best crops to plant are predicted is made in accordance with the environment of the time. This study provides a more accurate planting forecast. The evaluation of the crops from afar can be a major advantage of using technologies based on IoT for extensive farms. Depending on the type of plants in their farmed location on the aforementioned factors to improve the effectiveness of smart farming The forecast of cropping of data obtained from IoT sensors Using the methods, the prediction crop completed.[5]

With the increase in the need to solve various agricultural problems and to produce sustainable products from agriculture, technologies found their way into the agricultural sector. Vast amount of data acquired from these farms can be used decisively in effective crop manufacturing and harvest. Adopting these technologies into agriculture, also known as precision farming has shown favorable results in increased crop produce, cost reduction and establish a sustainable environment. Usage of modern technologies like Artificial Intelligence (AI), Internet of Things (IoT) and Blockchain are most efficient and advantageous in resolving the challenges faced in crop cultivation. Smart farming, vertical farming, and precision agriculture are some of the techniques developed from using previously mentioned technologies. Machine learning algorithms are utilized by taking several factors such as management of crops, production, pest, and disease control, etc., into consideration and are then analyzed and the predictions are made. Consequently, necessary changes can be made in the farm to facilitate more yield and profits to the farmers.[6]

Big Data Analytics, another latest technology can be integrated with different kinds of data i.e., data which is placed in a structural format and data which is unstructured. This data can either be collected from the farms with usage of sensors or from databases managed and regulated by various institutions. Big Data Analytics is a complete integration of two entities: 1) enormous datasets 2) aggregation of tools based on analytics consisting of different categories like statistics, predictive analysis, AI, data mining, and Natural Language Processing (NLP) which play a major role in business intelligence. Unlike conventional methods, these mechanisms can help in the early classification of pests and plant diseases, which is a major advantage. Deep learning algorithms like CNN have been incorporated more in the sensing of discrete applications remotely. The CNN model is capable of recognizing characteristics like leaf disease, estimation of crop, health and hygiene of the tree, height of the tree, species, canopy area, quality of the soil, etc. [7]

Cloud computing plays a crucial role in the maintenance of data and information related to web applications. This technology has the capacity to intensify the efficiency of data processing, minimization of expenditure, high security of information and flexibility of the usage of data. Applications using cloud computing serve as a potential key having low equitable cost, service costs and efficient usage of computational resources. Geographic Information Systems (GIS), a solution for farm management based on software facilitate in automation of analysis and collection of data, planning, management of farm, keeping records, making decisions and supervision of operations in an effortless manner. It is a tedious task for farmers to supervise complex agriculture related data in order to make fine decisions as a lot of field parameters are involved to manage a farm.[9]

Several challenges are faced while incorporating new-era technologies into the agricultural sector. Some of them are namely, collection of data, techniques in big data analytics, availability of infrastructure to facilitate computing, management of data and scalability of data in real-time scenarios and uncertainty of data management. Crop prediction approaches like Grey Wolf Optimization technique (GWO), Naïve Bayes, Apriori algorithm and K-means clustering are incorporated into Smart Farming where services like Mobile Computing, Cloud Computing and Internet of Things are discussed.[9]

Prediction of weather and climatic fluctuations also play an important role in making decisions related to agriculture. By incorporating Big Data Analytics techniques into agriculture, prediction can be made easily and effortlessly through automation. The model is built using a framework known as Hadoop to find a quick-fix to present day difficulties like food shortages due to the climatic changes, prediction of the impact due to extreme weather circumstances and the prevention of the aftereffects on global finance. Information and Communication Technology also known as ICT permits farmers to work as a team, setup association between them and share information dealing with crop predictions among them. As the farmers or the agricultural practitioners get associated with each other, various management frameworks defined by software are emerged. This can be very fortunate and handy to the peasants staying in rural precincts as a source of communication. Obstacles faced by these farmers can be overcome with the help of these eminent technologies. Some of these issues include limited availability of information regarding the crops, miniature farms, problems faced to ensure the quality of the crop farming and the gaps produced technically. As a result of resolving these arguments, the yield or the harvest of crops is improved significantly with the utilization of datasets based entirely on the weather conditions, crop details, soil compositions and plant diseases.[10]

Crops like rice, chili pepper, cotton, maize, wheat, soybean, etc., are studied and analyzed to predict the results based on the models developed. These models are built using various machine learning and deep learning algorithms like MapReduce, Logistic Regression, Naïve Bayes, Support Vector Machines (otherwise known as SVM), Artificial Neural Networks, etc. Data regarding the crops is collected from various meteorological stations, sensors, satellite images and ministries belonging to the agricultural sector. Several key features are taken into consideration for prediction analysis and in training of the models developed using deep learning and machine learning algorithms. These attributes include environmental factors like humidity, rainfall, average temperature, sunlight exposure and intensity, precipitation, air pressure, evaporation, and speed of the wind, etc. Linear Regression and MapReduce techniques produce far more accuracy than the other models in the forecasting of climate and weather. The characteristics of the usage of these techniques involve fast and accurate predictions, easy implementation, exhibits robustness towards noise in training datasets and high performance. Limitations include high consumption of power and computational resources, problems related to the overfitting of data, and adjustment of parameters is time intensive.[11]

Decision support systems based on Big Data for the selection of crops is proposed. These support systems mainly contribute in the monitoring of crop growth, quality production and management of the crop. Information regarding consumers, harvesting and yield of the crops, pests and disease data, data in the supply chain, soil related data and weather data, etc. All this environmental data collected is safely put in storage in the cloud. The cloud technology integrated with the Big Data Analytics tools make a big data-based decision support system. This data which is abstracted is used to train the machine learning models. Estimation of the amount of crop yield aims at examining features which produce a change and have a lasting influence on the production of the crop such as topography, composition of the soil, climatic conditions, crop diseases, and irrigation facilities, etc. For higher requirement estimation and productivity, these software tools can open doors for crops to be integrated within the universal supply chain. To support businesses in agriculture and the people in farm administration all around the world, the data dealing with real-time applications and circumstances are used. This information pertaining to events in real-life can help the farmers to make essentialpractical decisions.[12]

RELATED WORK:

All the states and their districts were included in the dataset named as crop production.csv. It includes information related to 125 crops, including details of the production of each crop and its geographic location. It was planted for six years, from 2000 to 2014 the Kharif, Rabi, summer, winter, and autumn seasons also the entire year. The main objective would be to map reduce to the data, then frame a Python recommender method for extracting output in accordance with the seasons and region. Later, perform k-means clustering by calculating the average yield per area, a team of A certain area will experience a crop's yield.

Big Data is nothing but information that is more diverse, which arrives at a faster rate and in larger amounts. Simply put, big data is more complex and huger set of data. Data mining includes preprocessing of data. A method for transforming unprocessed data into a format that is effective and very beneficial. Most of the cleansing of data, data integration, data transformation and data reduction. Data cleaning is the removal of redundant, incorrect, corrupted data from the dataset. Whereas, data integration is the merging of data from data sources that are heterogeneous into a consistent data storage. Data transformation is a mechanism which is done to join structured and unstructured data together for the sole purpose of analyzing it later and to find various patterns in them.

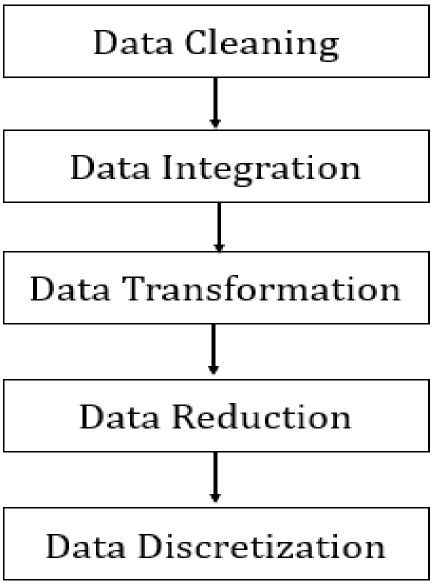


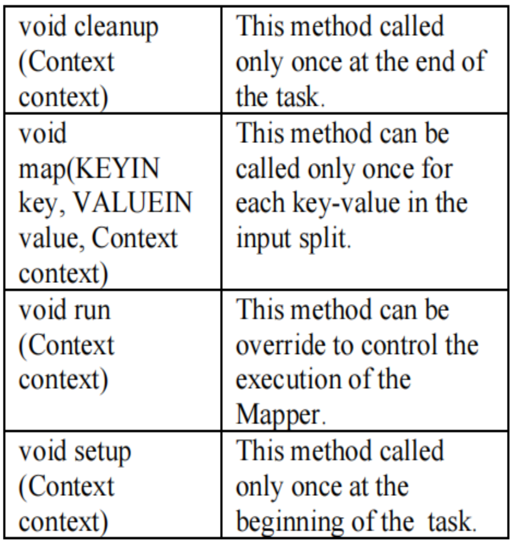
Fig 1: data pre-processing

A. Map Reduce:

MapReduce is a programmed model and processing method for distributed computing. A distributed computer system comprises of numerous software components present on multiple systems or computers. The computers can either be physical connected using a Local Area Network or systems which are located far away from each other geographically and can be connected remotely via a Wide Area Network (WAN). The MapReduce algorithm comprises of two crucial components. Specifically, Map and Reduce tasks. Map task requires a set a set of data into a different set of data, wherein individual components are divided into key-value pairs, or tuples.

Steps in MapReduce

1. The map accepts some pairs of data and gives the output list of "key and value" pairs. In this situation, keys will not be distinctive.
2. Using the results of the Map function, sort and shuffle the Hadoop architecture uses this kind of and these lists of "key, value" pairs are subject to shuffling and transmits a list of values and distinct keys. Connected to this special key, list(values)>.
3. Sending the reducer a sort-and-shuffle output phase. The reducer carries out a certain task. A table of values for the special keys, and Key and value output will be saved and shown below.

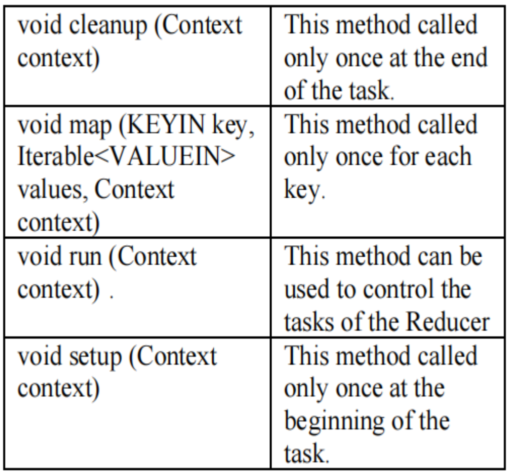


void cleanup (Context context)	This method called only once at the end of the task.
void map(KEYIN key, VALUEIN value, Context context)	This method can be called only once for each key-value in the input split.
void run (Context context)	This method can be override to control the execution of the Mapper.
void setup (Context context)	This method called only once at the beginning of the task.

Table-1: Methods in map1()

B. Map Function:

The output key-value pairs are produced by processing the incoming key-value pairs using the map function. It is possible that the map output and input types will differ. Year and region would be kept in the key in this paper, and the corresponding parameter for each month will be used as the result for such map function. To process the enormous volume of data, MapReduce is used. The forthcoming data must flow from multiple phases in order to be handled in a parallel and distributed format.



void cleanup (Context context)	This method called only once at the end of the task.
void map (KEYIN key, Iterable<VALUEIN> values, Context context)	This method called only once for each key.
void run (Context context) .	This method can be used to control the tasks of the Reducer
void setup (Context context)	This method called only once at the beginning of the task.

Table-2: Methods to reduce

C. Reduce Function

Each key which is distinct has the Reduce function applied to it. The order of these keys has already been determined. The corresponding values of the keys' can be used to iterate the Reduce function and provide the desired output. The production and area will be used as the value, and the region, the year, the season, and crop will serve as key.

D. K-Means Clustering

K-Means clustering is a type of unsupervised learning, which uses machine learning algorithms to evaluate and cluster unlabeled datasets, and clustering is a type of unsupervised learning algorithm. These algorithms employ continuous unlabeled data to cluster the dataset which is unlabeled into multiple clusters and identify underlying patterns or data grouping without the need for human interaction. Here, K specifies how many pre-defined clusters must be produced as part of the process; for example, if K=2, there will be two clusters, if K=3, there will be three clusters, and so on. It is an iterative approach which separates the unlabeled dataset into K distinct clusters, each of which contains just one dataset and shares a set of common characteristics within them. We are computing WCSS (within-cluster sum of squares) for each value of K; WCSS is the sum of squares calculating the distance between every point and the cluster's centroid, then drawing a chart between K and WCSS. The pronounced bend or sharp point of the graph. The plot has a point that resembles an arm, so that the optimal value of K is thought to be point.

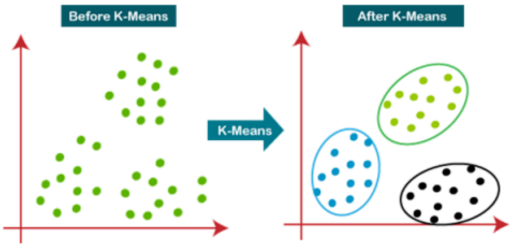


Fig 2: K- Means Clustering (Before and After)

Firstly, an elbow graph will be created in order to determine the ideal value of K and the total number of clusters needed. We will make use of the Scikit-learn to use a library. Afterward, by using the fit predict approach, one can obtain cluster values. An array will be used to do this. The values will be represented by integers beginning with 0. A single cluster. Plotting of the clusters will follow. Using the Matplotlib library's scatter technique. Each clustered centroid will be displayed, and thus indicate the average value generated by the cluster which would plot each crop, and each A distinct hue would be used to signify a cluster.

RESULTS AND DISCUSSIONS:

The state, region, and current month are used as inputs. It provides the season name, three crops that produce more crop, and seed based on the input. Additionally, it provides information on the temperature, wind direction, humidity, and water level, which increases yield based on the dataset. If Andhra Pradesh is the input, then July, Delta, and According to the feedback, rice is suggested. Banana, peppers If the, Rice could be displayed the approximate temperature is 80.6 and the actual rainfall is 1314.8 mm versus the anticipated 5 production area 7073335 Bananas might if the temperature is 80.7 degrees and the observed.

In terms of projected produce per area, rainfall is 2540.5 Variety should be Musa Parmigiana of 26.4545 if the soil is Musa and Garden soil (sand), Clay loam if the soil type is Indandamanensis. Expected wind speed is 5.2 and humidity is 79.0.

CONCLUSION

Using map reduce and K-means Clustering, this research introduces a crop recommendation system and provides an effective computation is the outcome. Based on the illustration K-Means clustering graphs, we can find the mean crop output for a variety of crops. Moreover, the relationship between crop characteristics and both 2D and 3D displays of the region graphs. The model concentrates on a variety of crops and their output by area, as well as the depending on the soil type and seed type types popular in a specific area. Eventually, it can be changed to become a crop recommendation. System alerts you when a sickness is present to that crop at that specific time of year.

REFERENCES:

[1]. Jiang, J., Zhang, Q., Yao, X., Tian, Y., Zhu, Y., Cao, W., & Cheng, T. (2020). HISTIF: A New Spatiotemporal Image Fusion Method for High-Resolution Monitoring of Crops at the Subfield Level. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 4607-4626.

[2]. Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. IEEE Access, 9, 63406-63439.

[3]. Elavarasan, D., & Vincent, P. D. (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. IEEE access, 8, 86886-86901.

[4]. Qiao, M., He, X., Cheng, X., Li, P., Luo, H., Tian, Z., & Guo, H. (2021). Exploiting Hierarchical Features for Crop Yield Prediction Based on 3-D Convolutional Neural Networks and Multikernel Gaussian Process. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 4476-4489.

[5]. Ashok, T., & Suresh Varma, P. (2020). Crop prediction based on environmental factors using machine learning ensemble algorithms. In Intelligent Computing and Innovation on Data Science (pp. 581-594). Springer, Singapore.

[6]. Gupta, R., Sharma, A. K., Garg, O., Modi, K., Kasim, S., Baharum, Z., ... & Mostafa, S. A. (2021). WB-CPI: Weather based crop prediction in India using big data analytics. IEEE Access, 9, 137869-137885.

[7]. Bendre, M. R., Thool, R. C., & Thool, V. R. (2015, September). Big data in precision agriculture: Weather forecasting for future farming. In 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 744-750). IEEE.

[8]. Majumdar, J., Naraseeyappa, S., & Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: application of big data. Journal of Big data, 4(1), 1-15.

[9]. Prediction Of Crop Yield In Precision Agriculture Using Machine Learning Methods. (2021).Webology.https://doi.org/10.29121/web/v18i4/122.

[10]. Channe, H., Kothari, S., & Kadam, D. (2015). Multidisciplinary model for smart agriculture using internet-of-things (IoT), sensors, cloud-computing, mobile-computing & big-data analysis. Int. J. Computer Technology & Applications, 6(3), 374-382.

[11]. Bhat, S. A., & Huang, N. F. (2021). Big data and ai revolution in precision agriculture: Survey and challenges. IEEE Access, 9, 110209-110222.

[12]. Rao, N. H. (2018, September). Big data and climate smart agriculture-status and implications for agricultural research and innovation in India. In Proc. Indian Natl. Sci. Acad (Vol. 84, No. 3, pp. 625-640).

[13]. Cravero, A., & Sepúlveda, S. (2021). Use and adaptations of machine learning in big data—Applications in real cases in agriculture. Electronics, 10(5), 552.

[14]. Ngo, V. M., Le-Khac, N. A., & Kechadi, M. (2019, June). Designing and implementing data warehouse for agricultural big data. In International Conference on Big Data (pp. 1-17). Springer, Cham.

[15]. Sahu, S., Chawla, M., & Khare, N. (2019). Viable Crop Prediction Scenario in BigData Using a Novel Approach. In Emerging Technologies in Data Mining and Information Security (pp. 165-177). Springer, Singapore.

[16]. Ranjani, J., Kalaiselvi, V. K. G., Sheela, A., & Janaki, G. (2021, December). Crop Yield Prediction Using Machine Learning Algorithm. In 2021 4th International Conference on Computing and Communications Technologies (ICCCT) (pp. 611-616). IEEE.

[17]. Sagana, C., Keerthika, P., Thangatamilan, M., Kamali, R., Nanthini, K., & Maghathani, S. (2022, January). Identification of Suitable Crop Based on Weather Condition. In 2022 International Conference on Computer Communication and Informatics (ICCCI) (pp. 01-06). IEEE.

[18]. Kumar, R., Singh, M. P., Kumar, P., & Singh, J. P. (2015, May). Crop Selection Method to maximize crop yield rate using machine learning technique. In 2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM) (pp. 138-145). IEEE.

[19]. Gupta, R., Sharma, A. K., Garg, O., Modi, K., Kasim, S., Baharum, Z., ... & Mostafa, S. A. (2021). WB-CPI: Weather based crop prediction in India using big data analytics. IEEE Access, 9, 137869-137885.

[20]. Bodapati, N., Himavaishnavi, J., Rohitha, V., Jagadeeswari, D. L., & Bhavana, P. (2022, March). Analyzing Crop Yield Using Machine Learning. In 2022 International Conference on Electronics and Renewable Systems (ICEARS) (pp. 1-8). IEEE.

[21]. S. Perla, N. N. K and S. Potta, "Implementation of Autonomous Cars using Machine Learning," 2022 International Conference on Edge Computing and Applications (ICECAA), 2022, pp. 1444- 1451.

[22]. Vineela, A., Lavanya Devi, G., Nelaturi, N., Dasavatara Yadav, G “A Comprehensive Study and Evaluation of Recommender Systems” Lecture Notes in Electrical Engineeringthis link is disabled, 2021, 655, pp. 45–53.