Revolutionizing Agriculture: A Comprehensive Analysis of Machine-Learning Techniques for Crop Prediction

[1]Anjali, [2]Kshiteesh R Bharadwaj, [3]Rohan Koundinya U H, [4]Varun R K, [5]Dr. Niharika P Kumar

[1,2,3,4]Student, Department of Information Science and Engineering,
[5]Associate Professor, Department of Information Science and Engineering,
[1,2,3,4,5]R V Institute of Technology and Management, Bangalore, Karnataka

Abstract - Research in agriculture is expanding. The cultivation of crops in agriculture depends significantly on environmental factors, including soil quality, temperature, humidity, and rainfall. Previously, farmers were able to exercise control over crop selection, monitoring of growth, and timing of harvest. However, modern-day farming faces difficulty in keeping up with rapid environmental changes. Consequently, outmoded prediction methods have been increasingly substituted by machine-learning techniques. Calculate agricultural production, frequent such approaches have been utilized in this regard. It is crucial to use effective feature selection techniques to transform the raw data into a Machine-Learning-friendly dataset to guarantee that a particular Machine-Learning (ML) model operates with high precision. Only data aspects that are significantly relevant in defining the model's final output should be included, which will decrease redundant data and improve the model's accuracy. To ensure that only the most significant structures are encompassed in the model, it is necessary to use optimum feature selection. Our model will become needlessly complex if we combine every characteristic from the raw data without first examining their function in the model-building process. Additionally, the time and space complexity of the ML model will rise with the addition of new characteristics that have minimal impact on the model's performance. The findings show that compared to the current classification approach, an ensemble technique delivers greater prediction accuracy.

Index Terms - Crop yield prediction, Random Forest.

I. INTRODUCTION

Numerous replicas have been created and tested as an outcome of the complicated process involved in predicting crops in agriculture. The utilization of numerous datasets is required since crop cultivation depends on both biotic and abiotic factors. "Biotic factors" refer to environmental elements that arise from the direct or indirect effects of living organisms on each other, including animals, parasites, predators, microorganisms, plants, and pests. In addition to natural factors, this category encompasses anthropogenic factors, such as irrigation, fertilization, plant protection, air and water pollution, and soil quality. These elements may guide variations in crop production, including internal defects, shape irregularities, and alterations in the chemical composition of the crop. Both Abiotic and Biotic factors influence the structure of the environment and affect the growth and quality of plants. Abiotic factors are written into 3 groups: chemical, physical, and other. The predictable physical factors are soil chemistry, salinity, soil type, atmosphere, geography, and soil type and rockiness. Furthermore, covered are climatic variables, ionizing, electromagnetic, ultraviolet, and infrared radiation, as well as mechanical vibrations (vibration, noise). Priority environmental pollutants comprise various chemical elements, such as sulfur dioxide and its byproducts, nitrogen oxides, and their byproducts, polycyclic aromatic hydrocarbons (PAHs), fluorine and its compounds, cadmium and its derivatives, nitrogen fertilizers, pesticides, and carbon monoxide. The others are asbestos, mercury, arsenic, dioxins and furans, and aflatoxins. Abiotic factors including bedrock, relief, climate, and water quality have an impact on it as3 well. Various soil-forming components affect how soils form in addition to how treasured they are for agriculture in different ways.

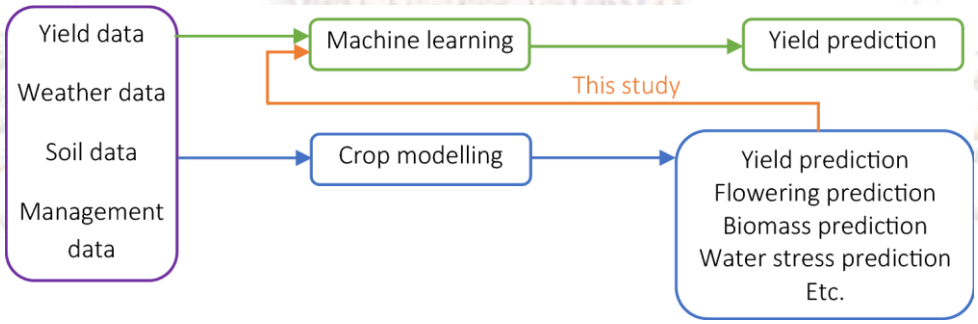

Fig.1.1 Crop Yield Prediction Block Diagram

Around a hundred different crops have been planted all over India. To facilitate comprehension and better visualization, these crops have been grouped. The Indian Government Repository provided the data for this study. With almost 2.5 lakh observations, the data consists of the following attributes: State, District, Crop, Season, Year, Area, and Production. To reduce error and produce more accurate predictions, we applied advanced regression techniques including Lasso, ENet, and Kernel Ridge. The format of this essay is as follows: a review of the literature, a methodology, a conclusion, and future work. Crop production prediction is neither straightforward nor easy. The strategy for predicting the extent of cultivation using statistical and mathematical methods is considered an iterative and enhancing optimization process, as per the findings of Myers et al. [5] and Muriithi [6]. It is also useful in the design, development, and formulation of new and improved goods. The presentation or conduct of statistical analysis necessitates the availability of numerical data. Based on

them, inferences about diverse occurrences are derived, and binding economic choices may be taken. According to Muriithis [6,] the further you represent occurrences numerically, the further you can say about them, and improving data quality allows you to receive more precise information and make more correct judgments.

II. LITERATURE SURVEY

A. Applying naive Bayes classification technique for classification of improved agricultural land soil: The advancements in computing and data storage have resulted in massive volumes of data. To bridge the knowledge gap, new techniques and methodologies were introduced such as data mining which resolves the difficulty faced during knowledge extraction from raw materials. In this study, the objective was to assess the success of newly developed data mining procedures and their potential application to a soil science database. The aim was to identify relevant associations that could be derived from the database. A large dataset of Soil database is extracted from the Department of Soil Sciences and Agricultural Chemistry, S V Agricultural College, Tirupati, the catalog holds dimensions of soil profile data from various locations of Chandragiri Mandal, Chittoor District.

B. Biotic components influencing the yield and quality of potato tubers: Throughout the last decade, potato yields in Canterbury have remained stable at around 60 t/ha. Potato growth representations, on the other hand, anticipate potential yields of up to 90 t/ha, which have previously been achieved by some commercial producers. A two-year study led by industry and research partners examined crop yield constraints. During the first growing season, 11 processing crops were thoroughly evaluated (final yield, plant health, and soil quality tests). Soil-borne infections (Rhizoctonia stem canker and Spongospora root infection), subsurface soil compaction, and inadequate irrigation management were identified as persistent drivers in lower yields.

C. Response surface methodology: A retrospective and literature survey: RSM is a grouping of statistical design and numerical optimization techniques used to optimize processes and product designs. The original research in this zone dates to the 1950s and has been extensively laboring, predominantly in the chemical and process industries. RSM has experienced widespread application and several innovative improvements over the last 15 years. This review focuses on RSM activity since 1989.

D. A Comprehensive Review of Crop Yield Prediction Using Machine-Learning Approaches with Special Emphasis on Palm Oil Yield Prediction: Crop yield prediction is a critical issue in the agricultural sector, and Machine-Learning algorithms have been increasingly used to estimate higher crop yields. This article provides a comprehensive analysis of the application of Machine-Learning algorithms in forecasting crop production, with a particular focus on palm oil yield prediction. The article begins by discussing the current state of palm oil yield around the world and explaining the commonly used features and prediction algorithms. It then presents a critical evaluation of cutting-edge Machine-Learning-based crop yield prediction, Machine-Learning applications in the palm oil business, and a comparative analysis of relevant works. The article also provides a detailed study of the advantages and difficulties related to Machine-Learning-based crop yield prediction and identifies current and future challenges in the agricultural industry. To address the present crop yield prediction issues, some remedies are suggested. The article emphasizes investigating the prospects of Machine-Learning-based palm oil yield prediction, covering areas such as remote sensing applications, plant growth and disease recognition, mapping and tree counting, optimum features, and algorithms. Finally, the article proposes a potential architecture for Machine-Learning-based palm oil yield prediction based on a critical review of existing related studies. This technology is expected to address new research issues in crop yield prediction analysis and develop an incredibly effective model for predicting palm oil yields with the least amount of computational difficulty.

E. Application of response surface methodology for optimization of potato tuber yield: The author studies the operational conditions required for optimal potato tuber yield production in Kenya. This will assist potato farmers in avoiding additional input costs. The potato manufacturing process was optimized using a factorial design 2 3 and a response surface approach. The combined impacts of water, Nitrogen, and Phosphorus mineral nutrients were explored and adjusted using the response surface approach. The best production parameters for potato tuber yield were determined to be 70.04% irrigation water, 124.75kg/ha of nitrogen provided as urea, and 191.04kg/ha of phosphorus supplied as triple super phosphate.

F. Crop yield prediction using Machine-Learning algorithm: Farming is the mainstay of the Indian economy, feeding more than half of the country's population. Machine-learning algorithms are cast-off to forecast agricultural production based on data such as rainfall, crop, and climatic conditions. Random Forest, the most popular and powerful supervised Machine-Learning method, can do classification and regression problems. Regardless of the distracting environment, they are utilized in crop selection to minimize agricultural yield output losses. Weather, climate, and other ecological aspects have wholly posed substantial threats to agriculture's long-term viability. Machine-Learning (ML) is significant because it provides a decision-support instrument for Crop Yield Prediction (CYP), which can assist with decisions such as selecting crops to farm and what to do during the crop's growing season. Crop yield estimation's main goal is to increase agricultural crop production, and it accomplishes this using a range of well-established models. Machine-Learning is becoming more popular around the world as a result of its success in a diversity of areas such as foretelling, fault recognition, pattern credit, and so on. Crop calculation is a critical agricultural topic. Farmers will be able to use the findings of this learning to govern the harvest of their crop before planting it in the agronomic field, allowing them to make educated decisions. To help farmers maximize agricultural yield, timely forecasting and analysis of future crop output are essential.

G. Using Hybrid Support Vector Regression to Predict Agriculture Output: Agriculture is not only a vital part of the rising economy but also necessary for survival. Crop production is difficult to predict since it is affected by numerous factors such as water, ultraviolet (UV), pesticides, fertilizers, and the extent of land covered in that region. This paper recommends II substitute Machine-Learning (ML) methods to analyze crop yield. These two techniques, Support-Vector-Regression (SVR) and Linear-Regression

(LR) are appropriate for estimating variable parameters in predicting continuous variable estimates with 140 data points obtained. The features labeled overhead are significant factors influencing crop yield. The fault percentage was designed via Mean Square Error (MSE) and Coefficient of Determination (R2), where MSE was roughly 0.005 and R2 was approximately 0.85. The same dataset was utilized to quickly compare the performance of the methods.

III. METHODOLOGY

In the temperate zone, it is difficult to determine exactly how different agroclimatic factors influence grain production during winter. To explore the effects of temperature and precipitation on wheat yield, we use data from the Indian Government. Indian database provides estimates of potential yields for cereals in different regions. It uses satellite images, crop models, and weather data to assess crop conditions and predict yields. The main factors that affect wintering yield are temperature days with temperatures above 5 degrees Celsius, frequency, and the number of days below 0. Temperature is one of the most significant factors in determining wheat yield. It can affect growth, development, and yield potential by affecting plant respiration rates, leaf area index, and water use efficiency. A higher temperature results in increased stomatal conductance (the ability of plants to captivate CO2 through their pores), which leads to faster photosynthesis rates. This means more sugars are produced leading to higher yields. Using publicly available data, many estimates can be made using regression models. One such prototypical has been cast off to assess the economic impact on grain producers and consumers if state intervention occurs in markets for agricultural commodities.

Agrometeorological parameters are needed to make accurate crop production forecasts, but some aspects of these components' fluctuation may constitute a special challenge. Several investigators have tried with varying degrees of success—to address this problem. The goal is to present a novel approach for modeling the effect of weather on crop manufacture via data from the Government sector. This approach relies on the use of linear-regression models, which are widely used in economics and other fields. The advantage of this method is that it allows researchers to assess how much variation in yield can be explained by variations in weather variable quantity without the creation of any norms around the underlying processes governing crop growth.

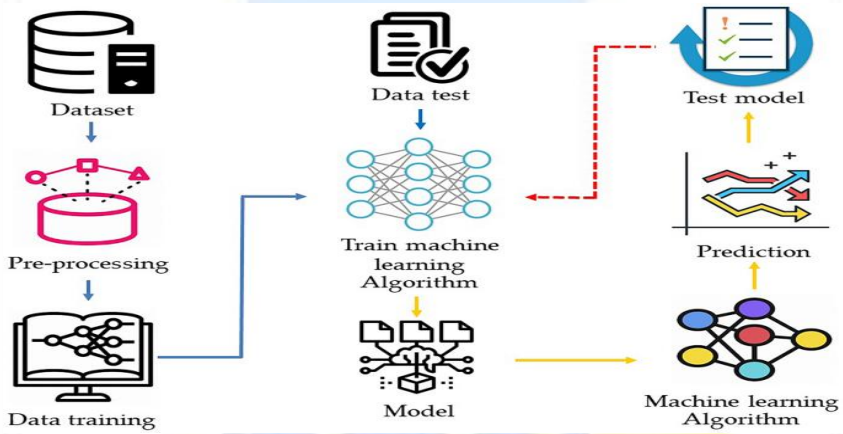

Fig.4.1 System Architecture

Disadvantages:

1. Because crop predictions are crucial to agriculture, and because soil- and environmental factors such as rainfall, humidity, and temperature affect these predictions so significantly— as well as the number of crops produced—it is significant that scientists account for these environmental factors when making their predictions.
2. Rapidly changing environmental conditions have made it difficult for farmers to continue farming as they use to do in the past. Reason of the unpredictability of weather, farmers have had strain constructing choices about when to plant their crops and what amount of water to use in irrigating them. The lack of information about the effects of weather variation on water resources has made it difficult for researchers to advance extra precise models that can be cast off by farmers when making decisions.

Advantages:

1. To avoid redundancies and enhance the precision of the ML model, only data characteristics which inflicted with a high degree of importance in deciding the model's final output should be included.
2. An ensemble method outperforms the previous classification technique in terms of prediction accuracy.

IV. MODULES

To carry out the aforementioned project, we created the modules listed below.

- Data exploration: The user will enter data into the system employing this module. Then the user will be able to create or delete datasets and visualize them.
- Data manipulation: this module contains tools for the manipulation of the data in our system. It allows us to perform batch operations, such as filtering and sorting, as well as transform tables into other formats (for example, from JSON to CSV).
- Processing: This module allows us to read data so that we can process it. We will be able to run SQL queries, which will allow us to extract the desired information from a dataset. We also practice this module to perform data analysis. Visualization: This module allows us to visualize our data in different formats (for example, as charts). It contains tools such as line or bar graphs and scatter plots so that we can easily identify trends in our datasets.
- We will be able to perform data analysis on the training dataset and visualize it using the visualization module. The test dataset, which remains untouched throughout this process, is then used to evaluate the recital of our model.

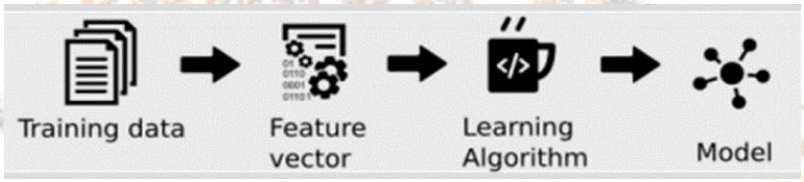

Fig. 5.1 Training Model

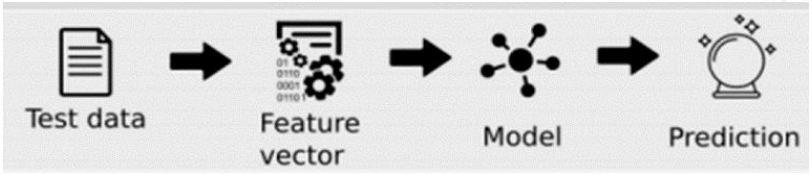

Fig. 5.2 Testing Model

- Model generation: Building the model with and without feature selection. - Feature Choice (SMOTE, ROSE, RFE, MRFE, BORUTA, MEMOTE) - Naive Bayes - KNN - Bagging Classifier - Random Forest Decision Tree - SVM - Gradient Boosting - Voting Classifier. Calculated algorithm accuracy.

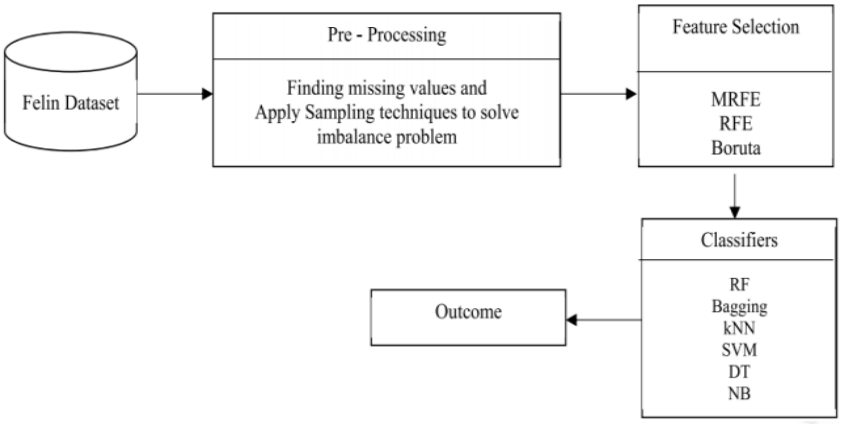

Fig. 5.3 Model Generation

- User Signup and Login: Using this module will create an account for a new user, who can then log in using the credentials provided during registration.
- User Input: By practicing this module user will give input and get an expected response.
- Predicted outcome: actual results. The user will be given a recommended response. Practicing this module will create an account for a new user, who can then log in using the credentials provided during registration. You can use it to generate users with different profiles and behaviors, or even different demographics.

V. PROPOSED A SYSTEM

The chief aspiration is to collect data that can be stored and evaluated to forecast crop yield. Machine-Learning methods are cast-off to predict agricultural productivity. This allows farmers to select the optimum crop for their needs. Goal to improve agriculture by attaining better results in crop yield estimation. A statistical model is built by practicing Machine-Learning methods and adequate optimizations to offer accurate and exact decisions. The results of this investigation will assist farmers in selecting the most suited crops to plant based on characteristics such as season and available acreage, with the least amount of risk of loss. The prediction process depends on the two fundamental techniques of feature selection [FS] and classification. Prior to the application of FS techniques, sampling techniques are applied to balance an imbalanced dataset.

Lastly, crop prediction issues are reviewed to define future directions. The processing unit on which the Machine-Learning model runs is a significant component, where it visualizes the data and improves the result.

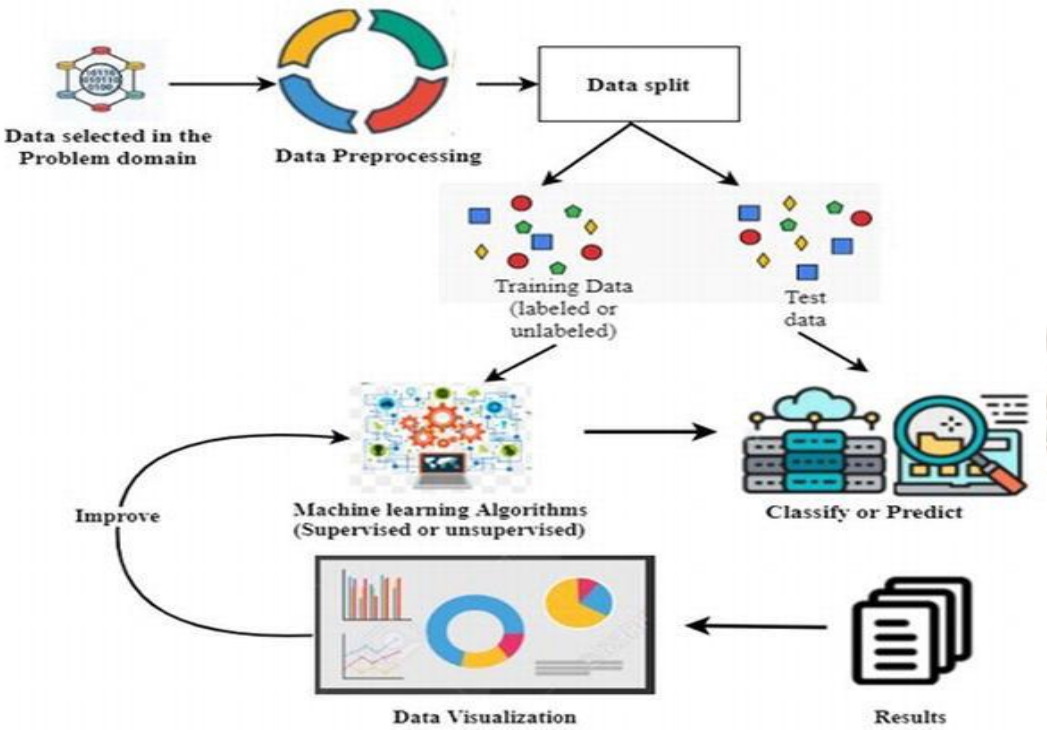

Fig. 6.1 Machine-Learning Approach

Advantages:

1. Only data features that are highly significant to defining the model's final output should be included to remove redundant data and increase the correctness of the ML model.

2. An ensemble technique outperforms the present classification approach in relation to prediction accuracy.

VI. CONCLUSIONS

The conclusion of this study focuses on the status of accurately predicting crop yields for effective agriculture management. The study employed different techniques to select the most relevant features and classify the crops to estimate their yields. The outcomes designate that the ensemble technique was the most accurate in predicting crop yields. These discoveries have noteworthy insinuations for the agricultural industry, as accurate crop yield predictions can assist farmers in making informed decisions about their planting strategies. The forecasts can be worn out on a local and national level for various crops, including grains, potatoes, and energy crops. Implementing these forecasting approaches can lead to substantial financial benefits for the agricultural industry. Overall, this study highlights the importance of accurate crop yield predictions and the potential benefits they can bring to the agricultural sector.

VII. FUTURE ENHANCEMENTS

According to the paper, the initiative has already achieved a wide range of its intended goals, and the next step is to progress the database for the system to store the extracted data. In adding together to this, there are several other areas where future research could be conducted, such as evaluating certain strategies in greater detail, exploring additional libraries, and investigating new approaches to methodologies. One area of interest for future research is the application of the Convolution model to real-world data. This could involve testing various plant aggregations and introducing more diverse plants into the procedure.

By doing so, it may be conceivable to advance a better identification of how different environmental factors impact yield. Another area that the researchers recommend exploring is the use of controllable environments for daily yield measurement needs. While the uncontrollable environment parameter helped to tool shed sunlit on the effects of environmental factors on yield, further study is needed to understand how this can be useful in practical situations. Finally, the researchers suggest developing an app for farmers to use and translating the entire system into different languages to make it more accessible to a wider audience. By doing so, it may be possible to

extend the reach of the initiative and help farmers around the world to optimize their crop yields and improve their overall agricultural practices.

VIII. REFERENCES

[1]. R. Jahan, ‘‘Applying naive Bayes classification technique for classification of improved agricultural land soils,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 6, no. 5, pp. 189–193, May 2018.
[2]. B. B. Sawicka and B. Krochmal-Marczak, ‘‘Biotic components influencing the yield and quality of potato tubers,’’ Herbalism, vol. 1, no. 3, pp. 125–136, 2019.
[3]. B. Sawicka, A. H. Noaema, and A. Gáowacka, ‘‘The predicting the size of the potato acreage as a raw material for bioethanol production,’’ in Alternative Energy Sources, B. Zdunek, M. Olszáwka, Eds. Lublin, Poland: Wydawnictwo Naukowe TYGIEL, 2021, pp. 158–172.
[4]. B. Sawicka, A. H. Noaema, T. S. Hameed, and B. Krochmal-Marczak, ‘‘Biotic and abiotic factors influencing on the environment and growth of plants,’’ (in Polish), in Proc. Bioróżnorodność Środowiska Znaczenie, Problemy, Wyzwania. Materiały Konferencyjne, Puławy, May 2017. [Online]. Available: https://bookcrossing.pl/ksiazka/321192
[5]. R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, ‘‘Response surface methodology: A retrospective and literature survey,’’ J. Qual. Technol., vol. 36, no. 1, pp. 53–77, Jan. 2014.
[6]. D. K. Muriithi, ‘‘Application of response surface methodology for optimization of potato tuber yield,’’ Amer. J. Theor. Appl. Statist., vol. 4, no. 4, pp. 300–304, 2015, doi: 10.11648/j.ajtas.20150404.20.
[7]. M. Marenych, O. Verevska, A. Kalinichenko, and M. Dacko, ‘‘Assessment of the impact of weather conditions on the yield of winter wheat in Ukraine in terms of regional,’’ Assoc. Agricult. Agribusiness Econ. Ann. Sci., vol. 16, no. 2, pp. 183–188, 2022.
[8]. J. R. Olędzki, ‘‘The report on the state of remote sensing in Poland in 2011–2014,’’ (in Polish), Remote Sens. Environ., vol. 53, no. 2, pp. 113–174, 2019.
[9]. K. Grabowska, A. Dymerska, K. Poáarska, and J. Grabowski, ‘‘Predicting of blue lupine yields based on the selected climate change scenarios,’’ Acta Agroph., vol. 23, no. 3, pp. 363–380, 2016.
[10]. D. Li, Y. Miao, S. K. Gupta, C. J. Rosen, F. Yuan, C. Wang, L. Wang, and Y. Huang, ‘‘Improving potato yield prediction by combining cultivar information and UAV remote sensing data using Machine-Learning,’’ Remote Sens., vol. 13, no. 16, p. 3322, Aug. 2021, doi: 10.3390/rs13163322.
[11]. S. H. Bhojani and N. Bhatt, ‘‘Wheat crop yield prediction using new activation functions in the neural network,’’ Neural Comput. Appl., vol. 32, pp. 1–11, Mar. 2020.
[12]. S. Khaki and L. Wang, ‘‘Crop yield prediction using deep neural networks,’’ Frontiers Plant Sci., vol. 10, p. 621, May 2019.
[13]. F. Abbas, H. Afzaal, A. A. Farooque, and S. Tang, ‘‘Crop yield prediction through proximal sensing and Machine-Learning algorithms,’’ Agronomy, vol. 10, no. 7, p. 1046, Jul. 2020
[14]. R. A. Schwalbert, T. Amado, G. Corassa, L. P. Pott, P. V. V. Prasad, and I. A. Ciampitti, ‘‘Satellite-based soybean yield forecast: Integrating Machine-Learning and weather data for improving crop yield prediction in southern Brazil,’’ Agricult. Forest Meteorol., vol. 284, Apr. 2020
[15]. Y.Jeevan Nagendra Kumar, V. Spandana, V.S.Vaishnavi, K.Neha, V.G.R.R. Devi, ‘‘Supervised Machine-Learning Approach for Crop Yield Prediction in Agriculture Sector,’’ IEEE 5th International conference ICCES, July. 2020.
[16]. Kevin Tom Thomas, Varsha S, Merin Mary Saji, Lisha Varghese, Er. Jinu Thomas” Crop Prediction Using Machine Learning” International Journal of Future Generation Communication and Networking Vol. 13, No. 3, (2020), pp. 1896–1901.
[17]. S. Shidnal, M.V. Latte, A. Kapoor “Crop yield prediction: two-tiered Machine-Learning model approach” International Journal of Information Technology. (2019), pp. 1-9.
[18]. E. Khosla, R. Dharavath, and R. Priya, “Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression,” Environ. Dev. Sustain., vol. 22, no. 6, pp. 5687–5708, Aug. 2020, doi: 10.1007/s10668-019-00445-x.
[19]. M Kalimuthu, P.Vaishnavi, M.Kishore “Crop Prediction using Machine-Learning” (ICSSIT2020) IEEE Xplore Part Number: CFP20P17-ART; ISBN: 978-1- 7281-5821-1.