International Journal of Data Science and Analysis
2020; 6(3): 72-82
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20200603.11
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Using Prescriptive Analytics for the Determination of Optimal Crop Yield

Terungwa Simon Yange[1, *], Charity Ojochogwu Egbunu[1], Malik Adeiza Rufai[2], Oluoha Onyekwere[3], Alao Abiodun Abdulrahman[2], Idris Abdulkadri[2]

[1]Department of Mathematics/Statistics/Computer Science, University of Agriculture, Makurdi, Nigeria

[2]Department of Computer Science, Federal University Lokoja, Lokoja, Nigeria

[3]Department of Computer Science, University of Nigeria, Nsukka, Nigeria

Email address:

lordesty2k7@ymail.com (T. S. Yange)

[*]Corresponding author

To cite this article:

Terungwa Simon Yange, Charity Ojochogwu Egbunu, Malik Adeiza Rufai, Oluoha Onyekwere, Alao Abiodun Abdulrahman, Idris Abdulkadri. Using Prescriptive Analytics for the Determination of Optimal Crop Yield. *International Journal of Data Science and Analysis*. Vol. x, No. x, 2020, pp. x-x. doi: 10.11648/j.ijdsa.20200603.11

Received: May 28, 2020; **Accepted**: June 8, 2020; **Published**: July 6, 2020

Abstract: The application of data mining has been utilized in different fields ranging from agriculture, finance, education, security, medicine, research etc. Data mining derives useful information from careful examination of data. In Nigeria, Agriculture plays a critical role in the economy as it provides high level of employment for many people. It is typical of farmers in Nigeria to plant crops without paying considerate attention to the soil and crop requirements as most farmers inherit the lands used for farming from their fathers and just continue in the pattern of farming they had always known. This is not favorable in the level of productivity they can actually attain as the effect can be seen in same level of crop yield year after year if not even worse. Modern farming techniques make use of data mining from previous data considering soil types, and other factors like weather and climatic conditions. This study built a model that enables possible prediction of crop yield from the historic data collected and offers suggestions to farmers on the right soil nutrients requirements that would enhance crop yield. This will enable early prediction of crop yield and prior idea to improve on the soil to increase productivity. The research used XGBoost algorithm for the crop yield prediction and the Support Vector Machine algorithm for the recommendation of appropriate improvement of soil nutrient requirements. The accuracy obtained for the prediction with XGBoost was 95.28%, while that obtained for the recommendation of fertilizer using SVM was 97.86%.

Keywords: Prescriptive Analytics, Optimal, Crop Yield, Machine Learning, Support Vector Machine, XGBoost

1. Introduction

The Nigeria agriculture is highly differentiated in terms of its climate, soil, water, crops, horticultural crops, plantation crops, medicinal crops, livestock, *etc.* Today, Agriculture contributes immensely to the Nigeria economy as 21.2 percent of Nigeria's GDP came from agriculture as at 2018 [1]. The agricultural sector is a major employer of labor but due attention has not been to it as Nigeria had relied on oil in the past decades to generate revenue and provide foreign exchange. The big drop in the price of oil in the past five years has exposed the relevance and interest in agriculture. According to the Federal Ministry of Agriculture and Rural Development in Nigeria, food (crop) production has not kept to pace with population growth, resulting in rising food imports and declining levels of national food self-sufficiency [2]. Nigeria is said to have in 2017 imported 23,192 tonnes of rice, 6,537 tonnes in 2018 and 2,380 tonnes in 2019 [2, 3]. However, with the recent ban of food importation in Nigeria, all hands need to be on deck to increase food production level as a nation; because with the high birth rate recorded annually, it will be challenging for local food supply to meet with the demand. Agriculture is facing the problem of changes in the resources that are directly affecting crop yield, so the agricultural productivities in Nigeria are unpredictable. For balanced and sustainable growth of agriculture, these

resources need to be evaluated, monitored and analyzed, so that proper methods can be put in place. Crop yield prediction is an important agricultural problem. Crop yield refers to the measure of seed or grains which is generated from a unit of land. This can also be referred to as both the measure of the yield of crop per unit area of land and the seed generation of the plant itself [4]. One of the metrics used in determining the efficiency of food production is crop yield. Simply, Crop yield is the amount of crop harvested per area of land. In the past, yield prediction was calculated by analyzing the farmer's previous experience on a particular crop in a particular piece of land [5]. There are several factors influencing the crop yield. These includes soil nature (which comprise of moisture, soil pH, soil temperature, soil type), the climate condition, and pest control; all these features are considered in predicting crop yield. Accurate information about the history of crop yield is an important thing for making decisions relating to agricultural risk management. Of all the factors considered in crop yield, we may not be able to control weather or climatic change but it is possible to enhance the fertility of the soil. Some researches concentrate on crop yield prediction while others concentrate on soil fertility recommendation but here we propose a system where crop yield can be predicted based on the nature of soil and a system that compares the nutrient strength of the soil and the required nutrients and recommends the appropriate fertilizer for optimal yield. This research is intended to bring enhancement to agriculture by achieving better results in crop yield in Kogi State with the use of machine learning tools which will help farmers to make good decisions in optimizing crop yield.

The remaining sections of this article is structured in the following order. Section 2 gives the survey of related works while Section 3 presents the basic materials and methods considered in this research. The implementation of system and results are presented in Section 4. In Section 5, the merits and demerits of results presented in Section 4 are discussed alongside the comparison between the results of this work and that of the existing works. As a conclusion, unique contributions of this article, limitations of the research and some future research directions are given in Section 6.

2. Review of Related Works

Prescriptive analytics is a type of data analytics in which the actions are determined as required in order to achieve a particular goal. This is a relatively new aspect of analytics that allows users to “prescribe” a number of different possible actions to guide them towards a solution. It uses optimization and simulation algorithms to advice on possible outcomes and answers. Thus, prescriptive analytics is all about providing advice [6]. Prescriptive analytics is able to suggest the best decision options in order to take advantage of the predicted future and illustrates the implications of each decision [7]. This attempt to quantify the effect of future decisions in order to advice on possible outcomes before the decision is actually made. It goes beyond descriptive and predictive analytics by recommending one or more possible cause (s) of action (s). The effectiveness of the prescriptions depends on how well these models incorporate a combination of structured and unstructured data, represent the domain under study and capture impacts of decisions being analyzed [7].

Technology as the fast-growing sector which has performed greatly in the agricultural sector by improving productivity. As a matter of fact, machine learning has emerged immensely on big data and high-performance computing to create new opportunities for data-intensive science in agro-technology domain. Thus, different types of machine learning technology are applied to solve agricultural sector problems most especially the yielding of crops. Fertilizer is one of the costliest inputs in agriculture and the use of the right amount of fertilizer is fundamental for farm profitability and environmental protection [8]. Location-specific fertilizer recommendations are possible for soils of varying fertility, resource condition of farmers and levels of targeted yield for similar soil classes and environment. Fertilizer Recommendation is very complex and dynamic in nature with the advancement of technology and with the progress of soil fertility. Also, Fertilizer Recommendation is based on the basis of soil test and crop response. Fertilizer recommendation systems such as the French system, Foster system, PORIM Open system and INFERS. These systems are based on leaf analysis, soil analysis, nutrient balanced approach, plant nutrient demand principles or their combinations. However, the Fertilizer recommendation system explicitly computes the nutrients required to correct the nutrient deficiency and meets the growth of the crops, and nutrient losses through environmental processes. The need for a critical search on how best to maximize the benefits from fertilizer inputs at economized rate would be rewarding.

2.1. Crop Yield Prediction Systems

Everingham *et al*. [9] predicted sugarcane yield using a random forest algorithm. they investigated the accuracy of random forests to explain annual variation in sugarcane productivity and the suitability of predictor variables generated from crop models. seasonal climate prediction indices and observed rainfall, maximum and minimum temperature, and radiation were supplied as inputs to a random forest classifier and a random forest regression model explained the annual variation in regional sugarcane yields. You *et al.,* [10] presented a deep learning framework for crop yield prediction using remote sensing data. Deep Gaussian Process were used for Crop Yield Prediction Based on remote sensor. It allows for real-time forecasting throughout the year and is applicable worldwide, especially for developing countries where field surveys are hard to conduct. They proposed a dimensionality reduction approach based on histograms and presented a Deep Gaussian Process framework that successfully removes spatially correlated errors, which might inspire other applications in remote sensing and computational sustainability. Monali *et al.* [11] presented a system where soil dataset was classified on the

basis of data mining techniques. The predicted category indicates the yielding of the crop. Naïve Bayes and k-Nearest Neighbor algorithms were used for crop yield prediction. Soil was categorized into high, low, medium category. Categorization of soil was utilized in the crop yield prediction. Rapid Miner tool was used in implementing the data mining technique. Hemageetha, *et al.* [12] mainly focused on the soil parameters like pH, Nitrogen, and moisture for crop yield prediction. Naive Bayes algorithm was used to classify the soil and 77% of accuracy was achieved. Apriori algorithm was used to associate the soil with the crops that could provide maximum yield. A comparison of accuracy achieved during classification using Naïve Bayes, J48 and JRIP was done and the best classification accuracy was used. Pandey & Mishra [13] carried out a crop yield prediction using climate parameters, Random Forest algorithm was used to train the model so as to get accurate prediction, 5 climatic parameters were chosen to train the model. The accuracy of the model: 77% which was found using 10- fold cross validation technique indicated a high correlation between the climate and the crop yield. Other agro-inputs such as soil quality, pest, chemicals used, *etc.* were not used as they change from field to field.

2.2. Recommendation Systems for Agriculture

Kuanr *et al.* [14] designed a collaborative recommender system that based on prior idea as regards the suitability of crops in specific location from the knowledge of previous months' weather condition gives recommendations to farmers. The system also recommends other seeds and pesticides depending on the location and farmers' preferences. The cosine similarity measure was used to determine similar user based on the location of the farmer. Prediction of the yield was by fuzzy logic and the system implementation was in Mamdani Fuzzy Inference model. Reddy & Kumar [15] in their recommendation system considers a set of conditions from which the suitability of an item can be predicted by providing suggestions for crops to be cultivated based on weather and soil; predictive modelling was applied from previous data collected and processed to determine the suitability of crops accordingly.

Raja *et al.* [16] saw the need of a system that can advise farmers on the appropriate crops to grow to meet demand. Non-linear regression technique was used to predict crop yield that a farmer can obtain from his land considering different factors like temperature, rainfall, past yield of crop, area of land and market prices. Pudulamar *et al.* [17] emphasized the need for crops to be chosen based on the requirements of the soil to increase productivity and they introduced precision agriculture utilizing K-Nearest Neighbour, Random tree, Naïve Bayes, CHAID techniques to build an ensemble model that will learn to recommend a crop under certain conditions efficiently. El-bendary *et al.* [18] proposed a recommender system at cultivation-time to predict best sowing dates for cereal crops during winter in Farafra Oasis. They believed that the ultimate utilization of farm resources will be supported by the system, calculating the growing degree days (GDD) fulfilment for each crop as well weather condition prediction, different machine learning algorithms were used in daily prediction of maximum air temperature from data of 25 growing seasons and using the M5P and IBk regression algorithms the performance far outweighed the other machine language algorithms that's been used in previous predictions based on the mean absolute error calculated and the best cultivation-time prediction was achieved by the M5P algorithm when tested with data for five growing seasons. Sangeeta & Shruthi [19] also emphasized on the low contribution of farming towards the GDP of India and suggested that prediction of best crop yield for certain regions considering atmospheric and land factors like rainfall, humidity, temperature, soil PH and soil type as well as the record of crops grown previously would have reasonable impact in agriculture. They proposed a system that recommends to farmers, crops that are suitable for their region with best yield. Viviliya & Vaidhehi [20] proposed a hybrid model which utilized the Naïve Bayes, J48 and association rules that recommends crops to states in southern India by putting into considerations geographic and climatic parameters. The essence of the system is to increase crop yield by the recommendation of suitable crop for their land. Lakshmi *et al*. [21] in their paper crop recommendation system for precision agriculture considered weather, geographical features, water utility and land and to build a simpler mechanism to predict the types of soil and the crops that are suitable to be grown in that soil. the system is in two parts. Map reduce weather is the first part used in data processing structure which calculate huge data set on a group of computers while nearest k neighbor technique was used between similar years which helped to calculate the weather distances. Madhusree *et al.* [22] designed a collaborative recommender system for the farmers in their work. The system gives prior idea regarding a crop which is suitable according to the location of the farmer based on weather condition of the previous months. The system also recommends other seeds, pesticides and instruments according to the preferences in farming and location of the farmers while purchasing the seeds through online. It used cosine similarity measure to find the similar user according the location of the farmer and fuzzy logic for predicting the yield of rice crop for Kharif season in state Odisha, India. The system was implemented in Mamdani Fuzzy Inference model. The results revealed that it provides prior idea regarding a crop before sowing of seeds. Mohmmad & Ali [23] in their work, a Survey on Agriculture Crop Recommender System discussed various approaches various approaches presented by different researcher on agriculture data analysis and the basic data mining approaches such as clustering, classification done by algorithms.

The farmers will be able to predict crop output before actual cultivation and also get necessary recommendations to improve the fertility of the soil so as to attain maximum output. such as K-means, SVM and PCA etc. They concluded that the agriculture sector contains dataset which is multidimensional and to deal with it, more algorithms are

needed to address the sector effectively. Veenadhari *et al.* [24] developed a software tool named Crop Advisor 'it is a user-friendly web page for predicting the influence of climatic parameters on the crop yields. C4.5 algorithms was used to find out the most influencing climatic parameter on the crop yields of selected crops in selected districts of Madhya Pradesh. The software provides an indication of relative influence of different climatic parameters on the crop yield. The accuracy of predictions was above 75 per cent in all the crops and districts selected in the study indicating higher accuracy of prediction.

Most existing researches focused on either prediction or recommendation. None has considered predicting the yield of a particular crop based on the soil nutrient in a particular location and at the same time recommending appropriate fertilizer for that same crop in that same location for optimum output. So it is paramount to have an all-in-one system to handle the prescriptive analysis of crop yield.

3. Materials and Methods

This described how the methods used in this research are deployed to achieve the purpose of this research.

3.1. Dataset Collection

The data used in this research was collected from Agricultural Development Programme (ADP), Ministry of Agriculture and the Nigeria Meteorological data station (NiMet), all in Lokoja, Kogi State, Nigeria. The data comprise of climate data, crop properties, humidity, rainfall, windspeed, area cultivated, latitude, longitude, temperature, pH, soil features and other crop yield features. The data about soil features include the following Calcium (Ca), Magnesium (Mg), Potassium (K), Sulphur (S), Nitrogen (N), Lime, Carbon (C), Phosphorus (P) and Moisture. For the purpose of this work, we focused on N, P, K on the side of the soil nutrient. A total of 79,345,678 records were collected for maize, sorghum, millet, yam, beans, pumpkin, spinach, rice, tomato, groundnut and cassava.

3.2. Proposed System

With the goal to improve the crop yield and also help the farmers in making good decision, we proposed a prescriptive model. The system has two basic features predicting and recommending the cropping practices. The prediction is based on past agricultural production, latitude, longitude, temperature and soil data. The prescription recommends good practice to achieve optimal crop yield. In this proposed system, we are applying data mining techniques on agriculture production-based datasets to find interesting pattern and trends in the data. With this, farmers would be able to know what happened in the past, predict what is going to happen and what should be done to get a better output. This is done in two phases. First, the prediction using the current state of the soil is done using XGBoost Algorithm. Secondly, the system uses SVM to recommend fertility improvement for the soil with an appropriate fertilizer. This is done by checking the soil type and nutrients, its reaction to the soil fertilizer given to the crops to achieve a recommendation for optimal yield of different crops in different region. The system recommends a fertilizer for a particular crop selected by the user/farmer using the collaborative filtering techniques that will be implemented with an SVM machine learning model to help check the accuracy of a recommendation of fertilizer for a crop. The schematic diagram for the system is shown in Figure 1. This consist of data collection (plant and soil analyses, and previous history), transformation and decision making.

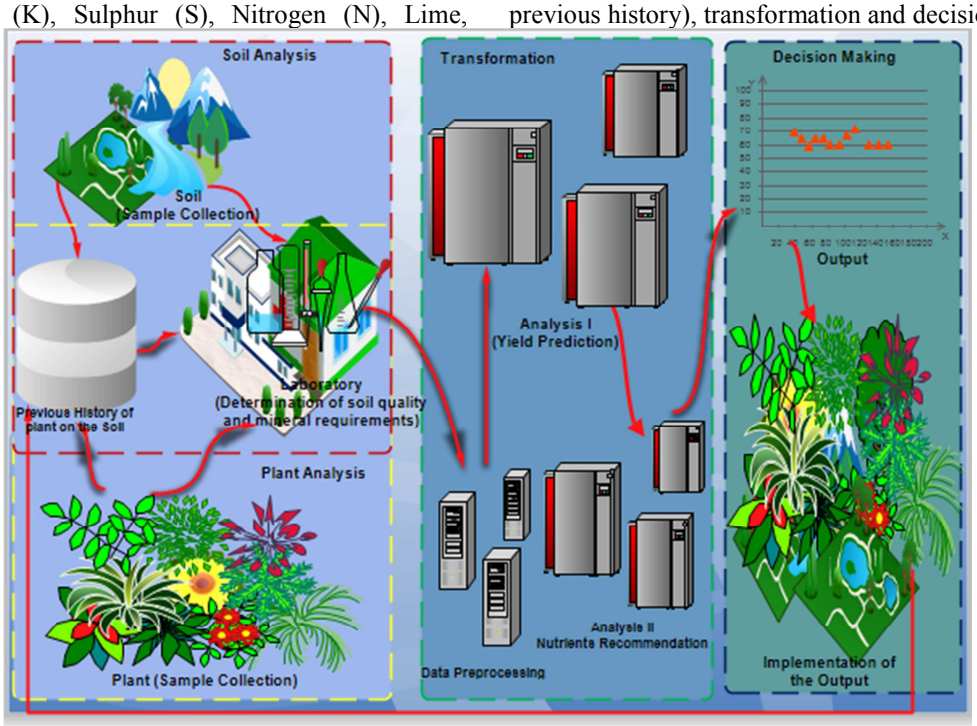

Figure 1. *Conceptual View of the Proposed System.*

Data Collection: This is the assessment of both the soil and the crop that is to be planted. For soil analysis, it involves fertility status assessment which the estimation of the available nutrient, determination of the amount of nutrient directly available in soil for subsequent uptake by crop, and determination of optimum fertilizer application ratio. This evaluate physio-chemical properties of the soil alongside with the environmental quality for the community hazards. For crop analysis, the nutritional status of crops and fertilizer needs. Mineral composition of plant is influenced by many factors which are also to be considered. Also, the previous history is considered. All these are done in laboratory and the result is forwarded to the next phase for proper interpretations and recommendations.

Transformation: Here, the processing of the data collected from the initial phase is done. The data undergo preprocessing which involve denoising, imputation, smoothing and normalization before it is ready for the analysis. The analysis is done in two (2) phases: prediction of yield considering the current state of the soil and recommendations for improvement of the soil fertility with appropriate fertilizer to achieve optimal yield. The classification of the soil and the prediction of the crop yield would be done by considering the soil and crop nutrient data, location and past history of the yield in the location. The XGBoost algorithm would be used to handle this task. Recommendation can be done using fertilizer data, crop and location data. In this part suitable crops and required fertilizer for each crop are recommended.

Decision Making: The output from the transformation phase are presented in visual form to the users in terms of charts. These are implemented by the farmers.

3.3. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a gradient boosting algorithm. In gradient tree boosting, updating weights of leaves of tree ensemble models, derivation an optimal model that can minimize the evaluation formula [25]. The derivation for the prediction as outlined in Maeda *et al.* [25] is as follows.

$$\gamma_i = \phi(x_i) = \sum_{k=1}^{n} K = f_k(x_i), f_k \in \mathcal{F} \quad (1)$$

Note that

$$\mathrm{F} = f(x) = wq(\mathrm{x})(q: \mathrm{R}m \rightarrow T, w \in \mathrm{R}T) \quad (2)$$

is regression tree space. x_i is input, γ_i is output, q tree structure, and T is number of leaves in tree. each f_k matches the independent tree structure q and the weight w. w_i represents the score of the ith leaf. This predicted value can be evaluated by

$$\mathrm{L}(\phi) = \sum_i l(\gamma_i, y_i) + \sum_k \Omega(f_k) \quad (3)$$

$$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda||\mathrm{w}||^2 \quad (4)$$

where l is the loss function to find the difference between the predicted value γ_i and the target value y_i.

Ω representing the complexity of the model is a regularization term and has the function of smoothing the weight to avoid over learning. This is adopted for this research.

3.4. Support Vector Machine

The Support Vector Machine (SVM) Algorithm is a non-linear generalisation of the generalised portrait algorithm developed in the 1960s, which is firmly grounded in the framework of the statistical learning theory. SVMs are linear learning machines, which mean that a linear function is always used to solve the regression problem. When dealing with non-linear regression, the input vector, x, is mapped into a high-dimensional feature space, z, via a non-linear mapping, and then conducting linear regression in this space [26]. The derivation given below is adopted from Du *et al.* [26].

Given a set of data points $G = \{(x_i, d_i)\}_{i=1}^{n}$ ($\mathrm{x_i}$ is the input vector, $\mathrm{d_i}$ is the desired value and n is the total number of data patterns), SVMs approximate the function using the following:

$$y = f(x) = w\phi(x) + b \quad (5)$$

Where ϕ (x) is the high-dimensional feature space, which is non-linearly mapped from the input space x. Coefficients w and b are estimated by minimising risk function R (C):

$$\text{Minimise R (C)}=\frac{1}{2}||\mathrm{w}||^2+\mathrm{C}\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(d_i, y_i) \quad (6)$$

Where

$$L_\varepsilon(d, y) = \begin{cases} |d-y| - \varepsilon & |d-y| \geq \epsilon \\ 0 & |d-y| \leq \epsilon \end{cases} \quad (7)$$

and ε is a prescribed parameter. The term $L_\varepsilon(d, y)$ is the so-called ε-insensitive loss function. This loss function defines a flat region which takes the flatness $y = f(x)$ as the centre, the thickness of which is 2ε. When the data samples are in the flat region, the loss is equal to 0, if the discrepancy between the predicted and the observed values is less than ". When the data samples are not in the flat region, linearity penalty is added to the function. The term $\frac{1}{2}||\mathrm{w}||^2$ is used as a measure of function flatness. The constant C, which influences a trade-off between an approximation error and the weights vector norm ||w||, is a design parameter chosen by the user.

To obtain the estimations of w and b, Equation (6) is transformed into Equation (8) by introducing positive slack variables ξ_i *and* ξ_i^*, as follows:

$$\text{Minimise R (w, } \xi^{(*)})=\frac{1}{2}||\mathrm{w}||^2+\mathrm{C}\sum_{i=1}^{n}(\xi_i, \xi_i^*) \quad (8)$$

$$\text{Subject to} \begin{cases} d_i - w\phi(x_i) - b \leq \mathcal{E} + \xi_i \\ w\phi(x_i) + b - d_i \leq \mathcal{E} + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

By introducing Lagrange multipliers α_i, α_i^* $(\alpha_i^*\alpha_i^* = 0, \alpha_i, \alpha_i^* \geq 0, i = 1, \ldots, n)$ and exploiting the optimality

constraints, we construct the Lagrange function according to the Lagrange dual theory.

$$L=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}(\xi_i,\xi_i^*)-\sum_{i=1}^{n}\alpha_i\,(\varepsilon+\xi_i-d_i+w\phi(x_i)+b)-\sum_{i=1}^{n}\alpha_i^*\,(\varepsilon+\xi_i^*+d_i-w\phi(x_i)-b)-\sum_{i=1}^{n}(\eta_i\xi_i+\eta_i^*\xi_i^*) \quad (9)$$

In Equation (5), η_i, η_i^* are the dual variables, which satisfy $\eta_i, \eta_i^* \geq 0$. The search for an optimal saddle point (w, b, α_i, α_i^*) is necessary because Lagrangian L must be minimised with respect to w and b.

As the optimal solution, we have

$$\delta_b L = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i) = 0$$

Specifically,

$$\delta_w L = w - \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi(x_i) = 0 \quad (10)$$

$$\delta_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

We obtain the dual problem by substituting (10) into (9). Specifically, the dual problem is as follows:

$$\max_{R(\alpha_i,\alpha_i^*)}: -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle\phi(x_i),\phi(x_j)\rangle-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{n}d_i\,(\alpha_i-\alpha_i^*) \quad (11)$$

Subject to $\begin{cases}\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)=0\\ \alpha_i,\alpha_i^*\epsilon[0,C]\end{cases}$ i=1,…, n

where $\langle .,. \rangle$ denotes the dot product in the feature space. Lagrange multipliers α_i, α_i^* are obtained by maximising function (11). Based on the nature of quadratic programming, only coefficients α_i, α_i^* will be assumed as non-zero, and the data points associated with them could be referred to as support vectors.

One basic idea in designing non-linear SVMs is to map input vectors x into vectors z of a higher dimensional feature space ($z = \phi(x)$, where ϕ represents a mapping). An input space (x-space) is spanned by components x_i of an input vector x, and a feature space (z-space) is spanned by components $\phi(x)$ of a vector z. By performing such a mapping, it is expected that the learning algorithm will be able to linearly separate images of x by applying the linear SVM formulation in a z-space. This approach is also expected to lead to the solution of a quadratic optimisation problem with inequality constraints in z-space. There are two basic problems in taking this approach: the first one is the choice of mapping $\phi_i(x)$; and the second problem is connected with a phenomenon called the 'curse of dimensionality'. This explosion in dimensionality can be avoided by noticing that training data appear only in the form of inner products $z_i^T z_j$, which are placed by inner products

$$z^T z_i = [\phi_1(x), \phi_2(x), \ldots \phi_n(x)][\phi_1(x_i), \phi_2(x_i), \ldots \phi_n(x_i)]^T$$

in a feature space; the latter is expressed by using the function:

$$K(x_i, x_j) = z_i^T z_j = \phi^T(x_i)\phi(x_j)$$

where $K(x_i, x_j)$ is named as the kernel function, which is a function in the input space. The basic advantage of using a kernel function lies in avoiding having to perform a mapping $\phi(x)$. Instead, the required inner products in a feature space are calculated directly by computing kernels for given training data vectors in an input space. Thus, using the chosen kernel, a SVM that operates in an infinite-dimensional space can be constructed. In addition, by applying kernels, one does not even have to know what the actual mapping $\phi(x)$ is.

The value of the kernel function is equal to the inner product of two vectors x_i *and* x_j , in feature spaces $\phi(x_i)$ *and* $\phi(x_j)$. That is, $K(x_i, x_j) = \langle\phi(x_i), \phi(x_j)\rangle = (\phi(x_i).\phi(x_j))$

Any function that satisfies the Mercer's condition can be used as the kernel function. It should be pointed out that training SVMs is equivalent to optimising Lagrange multipliers α_i, α_i^* with constraints based on function (11).

Then, the regression function given by Equation (5) has the following explicit form:

$$y = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle\phi(x), \phi(x_i)\rangle + b \quad (12)$$

$$y = f(x, \alpha_i, \alpha_i^*) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x, x_i) + b$$

With improving crop yield as the target, the recommendation of fertilizer for every purpose is done by checking the soil type and nutrients, its reaction to the soil fertilizer given to the crops and others techniques to achieve a better recommendation for different crops in different region. The proposed system would also improve novice farmers from recommending inappropriate fertilizer to crops. The application has two core features which include the detailed information (crop description documentation) and type of soil fertilizer for that particular crop in a particular region.

Fertilizer recommendation is emphasized to tackle the stated problems, during the fertilizer recommender system analysis with the five modules which include: crop information module, soil nutrient/test analysis module, fertilizer information module, and fertilizer recommendation module.

Crop information aspect of the system enables the selection of crops and viewing the information about the type of crop the user has intention to plant. In the area of the system, it is essential because it depends on the experts that will select or input the type of crop and not a novice who does not have knowledge about the system.

The fertilizer recommendation aspect of the system is a novel feature of this system for recommender system analysis.

Algorithms for Yield Prediction

The algorithm of the system shows step by step procedure

adopted by the system for easy operability. It describes how the user make requests on the system, and how the system responds to these requests.

Step 1: Start
Step 2: Load all XGBoost libraries
Step 3: Input dataset
Step 4: Data cleaning and feature engineering
Step 4: Split data in to training and testing
Step 6: Tune and run the model
Step 7: Fit a model to the training data
Step 10: Evaluate the train model using test data
Step 11: Display the result of the evaluation
Step 12: End

Algorithm for Fertilizer Recommendation

The algorithm of the system shows the stepwise process involved in the system for easy operability. It describes how the user makes requests to the system, and how the system responds to these requests. The steps below show how SVM algorithm is executed in a model (system).

1. Identify the right hyper-plane (condition-1): whereby given certain fertilizer for a particular crop, then it classifies the three fertilizer into three (3) or two (2) hyperplane, which are the Nitrogen, Phosphorus and Potassium and also select the hyper-plane which segregates the two classes accurately.
2. We identify the right hyper-plane thereby maximizing the distances between the nearest data point (either classes e.g. Nitrogen, Phosphorus, Potassium) NPK values therefore the right hyper-plane will help us to decide the right hyper-plane by checking the margin that is appropriate between the Nitrogen, Phosphorus and Potassium.
3. SVM selects the hyper-plane which classifies the classes Nitrogen Phosphorus and Potassium (NPK) values accurately which is prior to maximizing margin.
4. The maximize margin that occur between the hyper-plane for classified classes which are the Nitrogen, Phosphorus, Potassium (NPK) with the lowest distance will be selected.

4. Experiments and Results

The system was implemented using Python Programming Language. This technology has already made libraries to implement the algorithms used in this work. In developing the model, the dataset was split into two sets: training set and testing set. 70% of the dataset is used for training, while 30% of the dataset is used for testing. The training data was fit into the machine for training and the test set is used against the training set for validation. The accuracy was computed by analyzing how accurately a test set scores by learning from train set. Figure 2 shows the learning curve for the system. Figure 3 is the scatter plot showing the relation between the actual crop yield and the predicted yield, the models does well in prediction because the actual yield and predicted yield follow same trend. The dataset consists of some NaN and inconsistent value which output as outlier in the plot. The Figure 4 explains the importance of each feature used for prediction, area cultivated show as the most important feature in predicting the crop yield, while Figure 5 is a histogram showing the consistency in the crop yield.

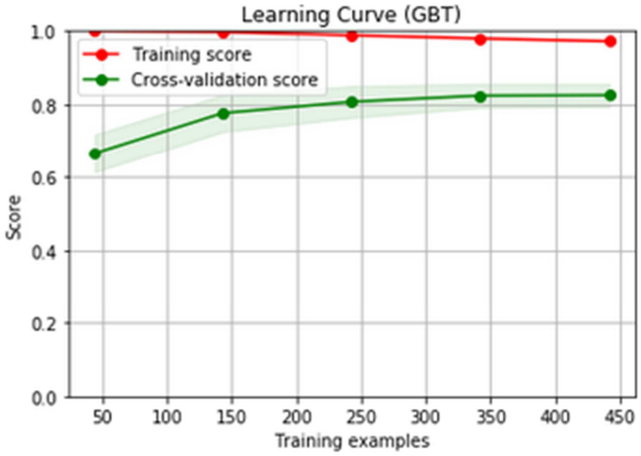

Figure 2. *Learning Curve.*

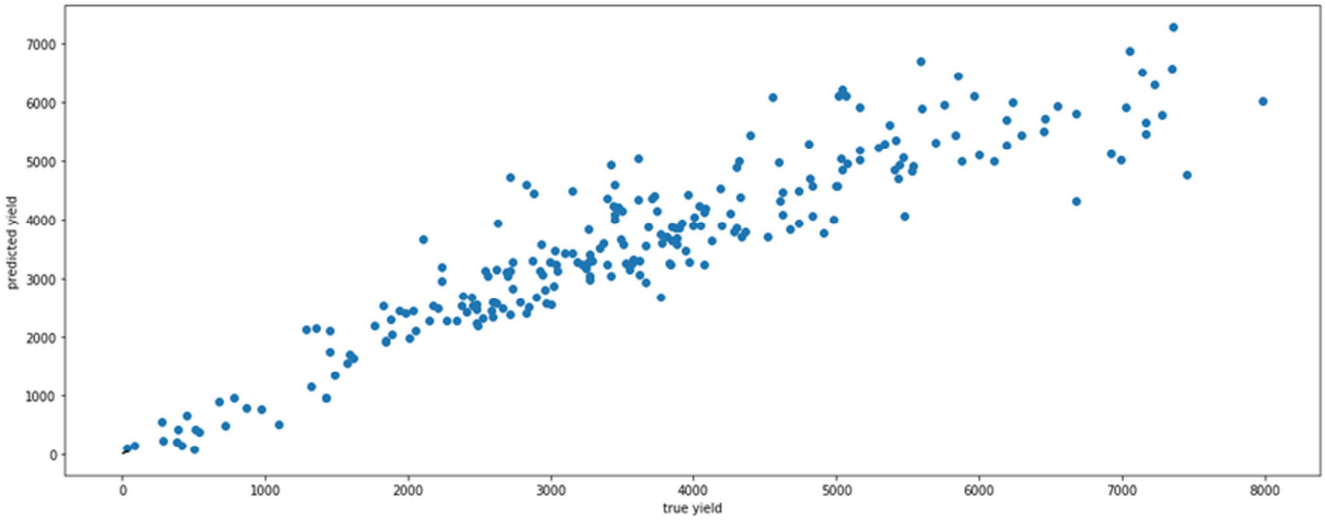

Figure 3. *Scatter plot of actual yield against predicted yield.*

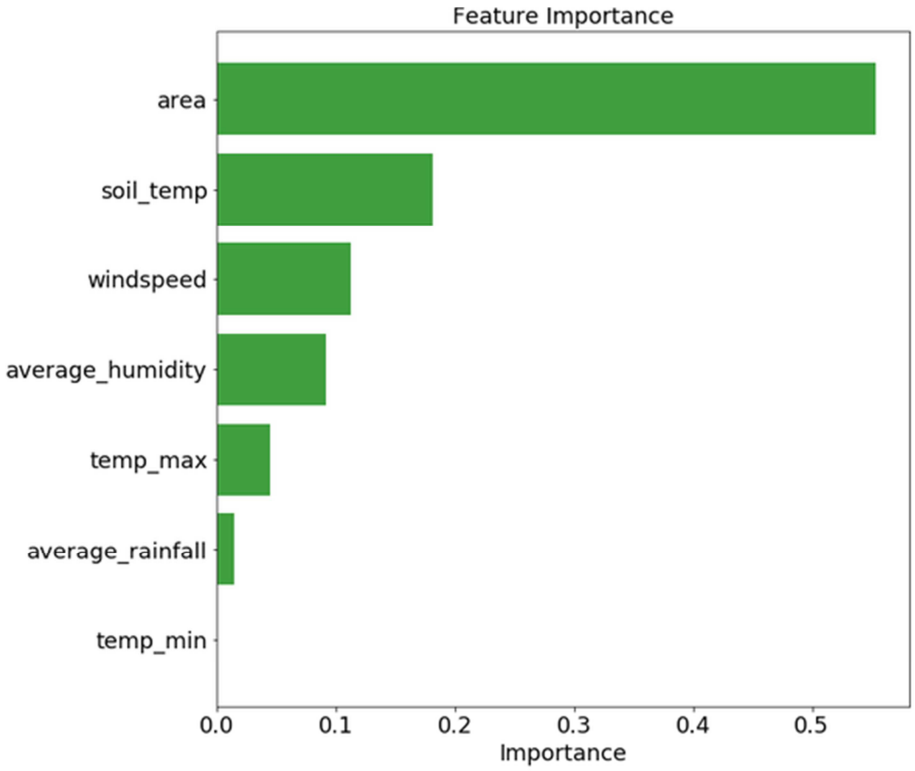

Figure 4. *Feature importance chart.*

Figure 6 shows the relationship of crop type which are represented as 1, 2, 3, 4 in the data which represent as vegetables, rice, tomato, and rice as the type of crop selected for the problem area. This explains the implication of soil nutrient with respect to their crop type. The graph is plotted on Nitrogen and Phosphorus respectively against the four type of classes of crops which are: rice cassava, spinach, pumpkin and sugarcane. This implies that given a certain quantity of fertilizer for a crop will be proportional to the same quantity of fertilizer given to another crop. Figure 7 described the reaction during the training of the dataset, whereby it describes the relationship of Loss with respect to the Number of epochs; the epochs describe the cycle to which the data complete its full training.

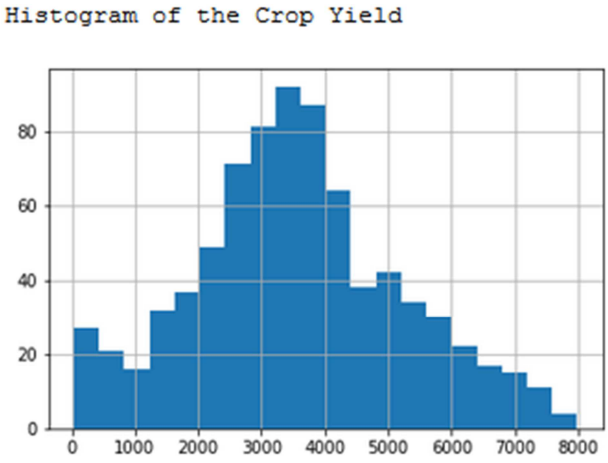

Figure 5. *Show the histogram of the crop yield.*

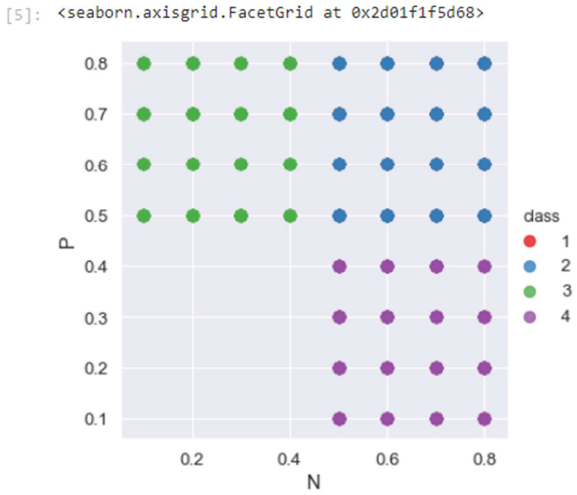

Figure 6. *Chart of the reaction of crop Type with the Fertilizer Quantity.*

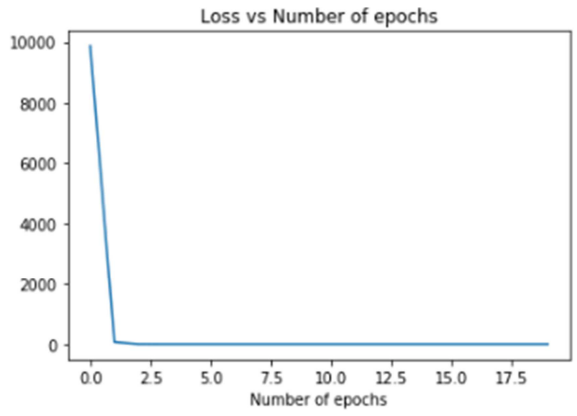

Figure 7. *Chart of relationship with the Loss to Number of epochs.*

Table 1 shows the predicted output and the predicted values for some crops using the current fertility of the soil using the XGBoost algorithm. The error in the prediction is also shown and the prediction has an average accuracy of 95.28%. Table 2 shows the nutrient requirements of the soil and the recommendation of the appropriate fertilizers that would improve the fertility of the soil, and predicted the yield that would be gotten if the fertilizers are applied. This was achieved with the SVM algorithm and the average accuracy was 97.86%. The error rate was also shown in the table.

Table 1. *Prediction.*

Crops	NF_Output	XGB_Err	XGB_Ac
Rice	1090.54	0.41841	99.58159
Cassava	6956.37	1.62610	98.37390
Spinach	4570.21	9.17431	90.82569
Pumpkin	23400.56	6.13718	93.86282
Sugarcane	4590.89	6.23671	93.76329

Table 2. *Recommendation.*

Crops	Nutrient Req	Rec_Fertilizer	NF_Output	WF_Output	SVM_Err	SVM_Ac
	N: 60%	Urea: 46-0-0				
Rice	P: 2%	NPK: 27-13-13	1090.54	1100000.12	2.32171	97.67829
	K: 5%	NPK: 27-5-5				
	N: 20%	NPK: 15-15-15				
Cassava	P: 60%	NPK: 12-12-17	6956.37	10956478.66	0.23475	99.76525
	K: 50%	NPK: 12-12-13				
	N: 40%	NPK: 3-1-2				
Spinach	P: 30%	NPK: 4-1-2	4570.21	6570023.78	1.38902	98.61098
	K: 50%	NPK: 3-2-2				
	N: 20%	NPK: 10-10-10				
Pumpkin	P: 60%	NPK: 20-20-20	23400.56	23230144.33	3.18791	96.81209
	K: 50%	NPK: 10-8-10				
	N: 60%	NPK: 19-19-19				
Sugarcane	P: 40%	NPK: 20-20-20	4590.89	4590166.45	3.54632	96.45368
	K: 55%	NPK: 20-19-19				

5. Discussion

The results presented in Section 4 are compared against other machine learning algorithms to justify our choices of algorithms. The algorithms used in the benchmarking include: linear regression, random forest, KNN and decision tree. The accuracy for prediction of the yield using linear regression, random forest, KNN and decision tree was 82.33%, 85.21%, 79.89% and 75.52% respectively. This when compared with 95.28% accuracy got using XGBoost algorithm, you would agree that the choice of the algorithm was right. For the recommendation of fertilizer and predicting of yield afterwards using linear regression, random forest, KNN and decision tree, we got 71.14%, 66.43%, 69.37% and 68.99% respectively. When these values are compared with 97.86% accuracy we got using SVM algorithm, you would agree that the choice of the algorithm was also right.

The research developed a system for prescriptive analytics. As discussed in Section 2 of this research, this involves prediction as well as recommendation of possible actions to be taken based on the predicted output. Prediction in this research was done by using the XGBoost algorithm. Most of the algorithms considered gave us fair accuracy, but the XGBoost produced the best accuracy of 95.28%. This was obtained using the different cultivation features which includes area cultivated, average rainfall, soil, temperature, average humidity, year, latitude, wind-speed and longitude. Comparing this result with that obtained by Madhusree *et al.* [22] which, the accuracy of the work was estimated at 74.4% when weather information was integrated at two intervals such as planting date to heading date and heading date to ripening date. Hemageetha, *et al.* [12] used soil parameters like pH, Nitrogen, and moisture for crop yield prediction. Naive Bayes algorithm was used to classify the soil and 77% of accuracy was achieved. The prediction result in this research outperforms these works considered. The result is as presented in Table 1.

To improve the yield of the crop, the phase first profiled the Soil using its nutrients data and then recommends the appropriates fertilizers that would augments the needed nutrients in order to get maximum yield. This is very essential for the identification of soil nutrient that is appropriate for crops; and also useful in identifying the soil nutrient, types of soils, and other properties effectively and accurately. The traditional recommender system includes the use of tables and flow-charts. This type of manual approach takes a lot of time, hence fast, automated system for fertilizer recommendation is needed to improve the yield of the crops. The fertilizer prediction is done using SVM with cross-validation where 70% data was used for training and 30% was used for testing and the error and accuracy were computed for all the crops considered (see Table 2).

6. Conclusion

The pivotal role agriculture plays in the Nigeria economy calls for intentional deliberations and advancements in increasing productivity. Technology pointed in this direction would enhance and complement the farmers' efforts to reach

this goal. Our work would help farmers to know in advance the deficiency in soil nutrient requirements and the appropriate fertilizer to boost soil competency for maximal crop yield. It will give farmers a prediction of expected crop yield and recommend appropriate soil nutrients that can increase crop yield prior before cultivation. Our future work will aim at large number of attributes with improved data set for specific crops comparing different climatic conditions and locations as well as more features of soil nutrients. It will also worth a future research to know the affluence of the fertilizer recommendation system on the soils' productivity over time.

References

[1] Plecher (2020). Distribution of gross domestic product (GDP) across economic sectors Nigeria 2018. Available at URL: https://www.statista.com/statistics/382311/nigeria-gdp-distribution-across-economic-sectors/.

[2] Federal Ministry of Agriculture and Rural Development (2008). Available at URL: https://fmard.gov.ng/.

[3] Ibirogba, F. (2019). Consumption of locally produced food items on the rise as borders remain shut. Available at URL: https://guardian.ng/saturday-magazine/consumption-of-locally-produced-food-items-on-the-rise-as-borders-remain-shut/.

[4] Veenadhari, S., Bharat, & M., Singh, D. (2018). Machine learning approach for forecasting crop yield based on climatic parameters. *International Research Journal of Engineering and Technology (IRJET),* 5 (3), 129.

[5] Priya, P., Muthaiah, U. & Balamuruga, M. (2018). Predicting Yield of the Crop Using Machine Learning Algorithm. *Ijesrt International Journal of Engineering Sciences & Research Technology,* 7 (4), 1-3.

[6] Bondre, D. A. & Mahagaonkar, S. (2019). Prediction of Crop Yield and Fertilizer Recommendation Using Machine Learning Algorithms. *International Journal of Engineering Applied Sciences and Technology,* 4 (5): 371-376.

[7] Alexandros, B. (2018) Prescriptive Analytics: A Survey of Approaches and Methods. Athens, Greece: National Technical University of Athens (NTUA).

[8] Kimetu, J., Lehmann, J., Ngoze, S., Mugendi, D., Kinyangi, J., Riha, S. V., Louis R., John & Pell, A. (2008). Reversibility of Soil Productivity Decline with Organic Matter of Differing Quality Along a Degradation Gradient. Ecosystems. 11. 10.1007/s10021-008-9154-z.

[9] Everingham, Y., Sexton, J., Skocaj, D. & Inman-Bamber, N.. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development. Vol 36 (2). 10.1007/s13593-016-0364-z.

[10] You J., Li X., Low M., Lobell D. & Ermon S. (2017). Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. *31th AAAI Conference on Artificial Intelligence* (AAAI 2017) Available at URL: https://cs.stanford.edu/~jiaxuan/files/Jiaxuan_AAAI17.pdf.

[11] Monali P., Santosh V. & Ashok V. (2015). Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach. *International Conference on Computational Intelligence and Communication Networks. 766-771. 10.1109/CICN.2015.156.*

[12] Hemageetha, N. (2016). A survey on application of data mining techniques to analyze the soil for agricultural purpose. *3rd International Conference on Computing for Sustainable Global Development (INDIA-Com)*, 3112-3117.

[13] Pandey, A. & Mishra, A. (2017). Application of artificial neural networks in yield prediction of potato crop. *Russian Agricultural Sciences*. 43. 266-272. 10.3103/S1068367417030028.

[14] Kuanr, M., Rath, B. K. & Mohanty, S. N. (2018), "Crop Recommender System for the Farmers using Mamdani Fuzzy Inference Model", International Journal of Engineering & Technology, 7 (4.15) 277-280.

[15] Reddy, K. A., & Kumar, K. A. (2018). Recommendation System a Collaborative Model for Agriculture. Available at URL: https://www.semanticscholar.org/paper/Recommendation-System-A-Collaborative-Model-for-Reddy-Kumar/173c7688bc6c86f6bdf122807a6e18f4ff9e0ae3.

[16] Raja S. K. S., Rishi R., Sundaresan E. & Srijit V. (2017). "Demand based crop recommender system for farmers," 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), Chennai, 2017, pp. 194-199.

[17] Pudumalar S., Ramanujam E., Rajashree R. H., Kavya C., Kiruthika & Nisha J. (2016). Crop recommendation system for precision agriculture. 2016 Eighth International Conference on Advanced Computing (ICoAC), Chennai, 2017, pp. 32-36.

[18] El-Bendary N., Elhariri E., Hazman M., Saleh S. M. & Hassanien A. E. (2016). Cultivation-time Recommender System Based on Climatic Conditions for Newly Reclaimed Lands in Egypt *Procedia Computer Science* Volume 96, Pages 110-119.

[19] Sangeeta & Shruthi G (2019) Survey on Crop Yield Recommender System in Agriculture. *International Journal of Scientific Research and Review* Volume 07, Issue 05.

[20] Viviliya B. & Vaidhehi V (2019). The Design of Hybrid Crop Recommendation System using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: Volume-9 Issue-2, 2278-3075.

[21] Lakshmi N., Priya M., Sahana S. & Manjunath C. R. (2018). CropRecommendation System for Precision Agriculture. *International Journal for Research in Applied Science & Engineering.* 6 (5). http://doi.org/10.22214/ijraset.2018.5183.

[22] Madhusree, K., Bikram, K. R. & Sachi, N. M. (2018). Crop Recommender System for the Farmers using Mamdani Fuzzy Inference Model. International Journal of Engineering & Technology, 7 (4.15) 277-280.

[23] Mohmmad S. & Ali, A. M. (2018) A Survey on Agriculture Crop Recommender System. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 7 (2).

[24] Veenadhari, S., Misra, B. & Singh, C. D. (2014). Machine learning approach for forecasting crop yield based on climatic parameters. *International Conference on Computer Communication and Informatics*, 1-5.

[25] Maeda, Y., Goyodani, T., Nishiuchi, S. & Kita, E. (2018). Yield Prediction of Paddy Rice with Machine Learning. In Proceeding of *Int'l Conf. Par. and Dist. Proc. Tech. and Appl. (PDPTA'18),* 361-365.

[26] Du, X., F., Leung, S. C. H., Zhang, J. L. & Lai, K. K. (2013). Demand Forecasting of Perishable Farm Products using Support Vector Machine. *International Journal of Systems Science,* 44 (3): 556-567.