JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-15, No.-5, May (2020) pp 90-100 ISSN (Print) 0973-8975

A NEW CROP YIELD PREDICTION SYSTEM USING RANDOM FOREST COMBINED WITH LEAST SQUARES SUPPORT VECTOR MACHINE

R. Mythili[1], AdityaVenkatakrishnan[2], T. Srinivasan[3], P. YashwanthSai Kumar[4]

[1]Assistant Professor, Department of Information Technology, SRM Institute of Science and Technology, Ramapuram, Tamilnadu, India.

[2,3,4]UG Students, Department of Information Technology, SRM Institute of Science and Technology, Ramapuram, Tamilnadu, India.

[1]mythilir2@srmist.edu.in, [2]nvkaditya@gmail.com, [3]srini125vazz@gmail.com, [4]yashwanth.saikumar14@gmail.com

Corresponding Author: R. Mythili

https://doi.org/10.26782/jmcms.2020.05.00008

Abstract

Predominantly in India, Agriculture is the most significant income generating segments and also a wellspring of endurance. Various occasional, financial and natural incidents impact the yield creation, yet erratic changes in these cases lead to an incredible misfortune for the Farmers. These dangers are to be decreased by utilizing reasonable mining methodologies on the identified data of soil type, temperature, environmental weights, mugginess and yield type. While, harvest and climate gauging can be anticipated by getting valuable bits of knowledge from this agricultural information that guides the Farmers to choose the yield, meanwhile they may need to plant for the expected year prompting extreme benefits. This paper presents an overview of different calculations utilized for climate, crop yield, and harvest forecast of the proposed crop yield prediction method using Least Squares Support Vector Machine (LS-SVM).

Keywords: Crop yield prediction, Support Vector Machine, Least Squares Support Vector machine, Data Analytics, Agriculture.

I. Introduction

In India, Agriculture is one of the important industrial sectors. Rural sustainability is the major factor on which the country's economy is highly dependent. Climate changes, unpredicted rainfall, decrease in water level, excessive pesticide usages etc., are the major factor based on which the growth level of Indian agriculture is decreased. Hence the crop production level can be effectively increased through the descriptive agriculture data analytics [XIII]. The proposed system mainly

aims at introducing an effective methodology on descriptive crop yield production data analytics [III], [XVII]. Although, there are some well-known statistical studies about Indian agriculture [XI] on crop yield prediction with soil behaviour[I], [XIV], [XVI], historic climatic & production data, it is necessary to manipulate the latest technology to solve the issues. Artificial Neural Networks (ANN) [VI] is suggested for such data analytics of classification, clustering, quantization, arrangements, approximation, forecasting, ascendancy and optimization. In the proposed system, crop yield classification is done to categorize crops on the basis of yield productivity. The class labels used are low, mid, and high. And regression [VII], [VIII] is performed to get the estimated crop yield cost. Yield forecast models are prepared on the basis of weather studies for estimating crop yield much before the actual harvest. Using correlation and regression techniques through empirical statistical models, crops yields are forecasted on an operational basis for the country. The proposed system uses the data analytics for predicting various meteorological parameters at various crop growth stages. In future, the proposed crop yield prediction system could be a complete recommender system for the Farmers.

II. Existing Systems

Most of the existing crop prediction frameworks [V] deal with the information that enables the farmer to get into the understanding of various yield expenses in the market. Various information such as mining calculations [XV], for example, Naive Bayes classifier, J48, K-Means were utilized in those frameworks. Similarly, it gives the soil characterization depending on Naive Bayes, Genetic calculation, Association Rule Mining. Clustering uses the data of soil database. This section consolidates the major crop prediction related works and the findings.

AnupamaMahato et al. [II]analyses thecomprehensive loss rates of rainstorm, storm hails and cold and flood disasters. Also discussed about the increasing trend of rates 0.45%, 0.29% and 0.72% respectively.

JapneetKaur et al. [X] described the food productivity as the important issue of developing countries like India. A new Parallel Layer Regression (PLR) along with Deep Belief Network (DBN) strategy is developed for performing the crop productivity estimation and implemented with five specific crops. The experimental results proved that accuracy (ACC), sensitivity (SEN) and specificity (SPE) are essential for accurate crop productivity prediction.

Pratap S. Birthal et al. [IV] has stressed the power of geometrics and retrieves the synoptic and substantial changes in cropping pattern. They also achieved the contemporaneous cropping system evaluation of yield parameter scrutiny for major crops like rice, wheat, sugarcane, and onion. And the test results are demonstrated a correlation r2 value of 0.834 with the estimated crop yield and normalized vegetation index. In their work, pH, electrical conductivity, and organic carbon of the soil are the surveyed information which were used to examine the spatial discrepancies of rice-based cropping system's productivity using spatial interpolation.

G. P. Zhang et al. [XIX] implemented a very simple system to remove any possible systematic component if left at all, by using an Artificial Neural Network

(ANN) which intends to filter out the residuals from a multivariate time series causality model.

Iooss et al. [IX] in his system justifies the global sensitivity analysis (GSA) of soil parameters using variance-based method. The system shows that the higher the criteria, better the parameter estimation quality.

Christian Baron et al. [XII] exploited remote sensing data for assessing crop productivity by building various indices such as Vegetation Condition Index (VCI), Temperature Condition Index (TCI) and Normalized Difference Vegetation Index (NDVI). They proposed a novel framework namely eXtensible Crop Yield Prediction Framework (XCYPF). The system has the provision towards precision agriculture for crop selection, dependent & independent variables, crop yield prediction datasets.

III. Proposed System

The proposed framework essentially focuses on climate estimating, crop yield expectation and harvest cost gauging. These elements help the farmers to develop the best nourishment harvests and raise the correct animals with understanding to natural parts. Also, the farmers can adjust to atmosphere changes somewhat by moving planting dates, choosing assortments with various development span, or changing yield pivots. For test investigation, the factual numeric information identified with horticulture is attempted. Though, the grouping-based strategies and regulated calculations are used for dealing with the gathered factual data. The arrangement techniques that are utilized in the framework are (1) Random Forest (RF) (2) Support Vector Machine (SVM) (3) Logistic Regression (LR) (4) Neural Network.

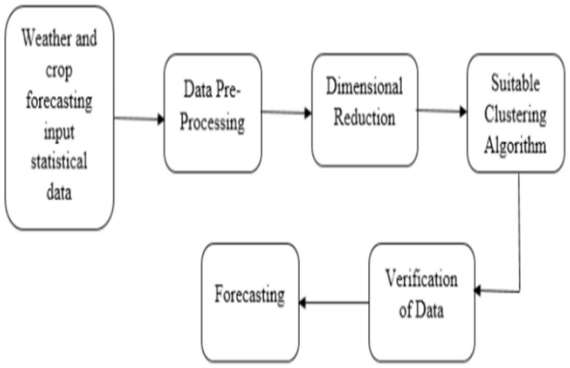


Fig.1: General flow of Proposed System.

These strategies will help in anticipating the precipitation, crop yield gauging and cost expectation of crops. Accurate data about history of harvest yield is something imperative for settling on choices identified with rural hazard the board. Along these lines, this paper proposes a plan to foresee the yield of the crop. The farmer will check the yield of the harvest according to the section of land, before developing onto the field.

IV. Methodology

Data Mining is widely applied to agricultural issues. Data Mining is used to analyse large data sets and establish useful classifications and patters in the data sets. Data mining [XV] process has the goal to extract the information from a data set and transform it into understandable structure for further use.Theproposed system analyses the crop yield production based on available data. The Data mining technique was used to predict the crop yield for maximizing the crop productivity. Figure. 1.shows the flow of proposed crop yield prediction.

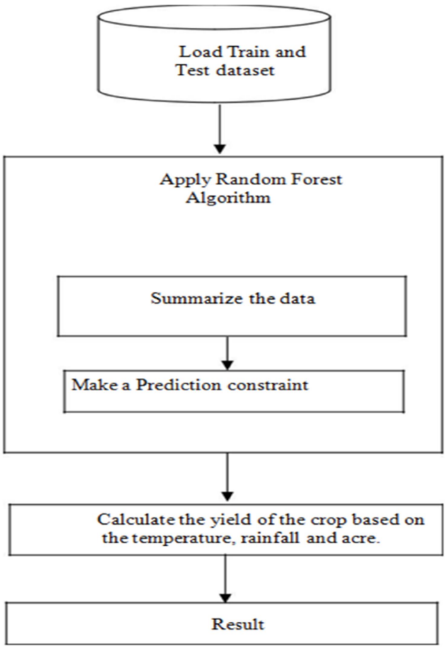


Fig.2: System Work Flow.

Data Pre-processing: Data Pre-processing organizes the selected data by formatting, cleaning and sampling from it. Below-mentioned three are the common data pre-processing steps:

Formatting- The data you have selected may not be in a format that is suitable for you to work with. The data could also be in an electronic database and you would like it during a file, or the info could also be during a proprietary file format and you would like it during an electronic database or a document.

Cleaning-Cleaning data is the deleting or fixing of missing or empty data. There could also be data instances that are incomplete and don't carry the info you think you would like to deal with the matter. These instances may need to be removed. Additionally, there could also be sensitive information in a number of the attributes and these attributes may have to be anonymised or removed from the info entirely.

Sampling- There may be far more selected dataset available than that needs to be worked with. More data may result in for much longer running times for algorithms and bigger computational and memory requirements. It can be taken a smaller stratified sample of the chosen data which will be much faster for exploring and prototyping solutions before considering the entire dataset.

V. Modules

Data Source

The creator proposes the approach as follows, taking an information Training Dataset, for example, Mushroom or Soybean [XVIII]. PSO-SVM Feature Selection Algorithm is applied for the choice of significant features from the dataset based on which classification should be possible precisely. This examination prescribes for approach creators to settle on proactive choices in recognizing which variables are the most critical to expand profitability. From the outcome, three significant things are concluded.

(1) Out of all qualities utilized, compost use has the most elevated prescient force.
(2) Out of the three calculations tried, J48 has demonstrated increasingly prescient force. Obviously, the entirety of the three calculations have demonstrated nearly a similar efficiency.
(3) The information might not have efficient unsurprising force as just a single year.

Moisture	rainfall	Average Humidity	Mean Temp	max Temp	Min temp
12.80168453	0.0123605	57	62	71	52
12.85165378	0.00417157	57	58	73	43
12.7767735	0	56	58	69	46
12.94200101	0.03174683	62	56	70	43
12.98465248	0	65	56	70	42
12.96447065	0.02719149	65	58	70	46
12.73799817	0.02682104	61	56	70	42
12.81938179	0.01028368	58	57	72	42
12.88390946	0.02046472	63	60	76	45
12.78451285	0.06005408	62	59	71	47
12.96881184	0.0841193	56	58	69	46
12.7843595	0	63	56	70	42
12.94458595	0	67	58	72	43
12.92528314	0.12447929	58	60	75	46
12.82106565	0.07450455	59	58	68	49
12.93921935	0.09858411	53	60	68	53
12.81587164	0.2228459	62	60	69	50
13.05680988	0.12855534	59	56	67	45
12.89795679	0.11341892	58	54	68	40
13.0346369	0.08006364	58	53	68	38
13.02395422	0.0776757	66	55	63	47
12.86211512	0.09419903	71	57	68	46
12.85579057	0.08925248	72	53	63	43
12.93984685	0.11055694	65	53	66	40
12.94620194	0.12220198	84	55	60	50
12.83171628	0.1796935	73	56	69	43

Fig.3: Dataset.

Clustering

Clustering is where the classifications properties which are related into different subsets, with the end goal that the crossing point of any two subsets brings about an invalid set. This goes under a solo issue where the information is unlabelled and not nonstop, this clustering system is broadly utilized for gathering the information dependent on the classifications. For this present issue of examination of the horticultural information, this gave us perhaps the best outcome as the connection of traits between each is comprehensively perceived. Regarding properties like normal temperature, precipitation we can aggregate this utilizing the clustering methods so the expectation model will be more exact than the directed preparing. In view of the anticipated properties the clustering systems can be applied to two sorts. Initially two-dimensional clustering where the two qualities are taken as information and order them into groups and three-dimensional bunches where n properties are taken and characterized. Since the information has barely have any characteristics and more exactness is normal, while picking a two-measure clustering model. With regards to preparing the precipitation informational collection, the time unpredictability is expanded as the gathering take more calculation power than the typical straight model. Tested procedures on the dataset utilizing clustering gave better outcomes of more precision than different models however the time taken by the calculation is more.

Bayesian Network

A Bayesian network is a probabilistic graphical model which can be utilized for measurable examination of the properties for a given dataset. Right now, traits are spoken to in graphs which are coordinated hub by hub. The hubs speak to the probabilistic capacity, and the edges speak to the restrictive conditions of the characteristics. The factually fit capacities can be determined outdrawn from the given graphical models where the expectations can be made. This methodology is especially helpful for natural demonstrating on the grounds that foreseen examples may rise at an assortment of scales, compelling a variety of model structures. Right now, didn't prepare a Bayesian network however a basic chart is spoken to how we would future be able to utilize it for an expectation model. This methodology unequivocally offers with the vulnerability of realities and connections and may comprise of both subjective and quantitative variable. The drawbacks of the Bayesian network is that it can't be applied to the huge datasets and hardly any capacities set aside more calculation effort for the expectation. Not many of the capacities which can be utilized on the agrarian information are sigma capacities and cross-corpora capacities.

Artificial Neural Networks

Artificial Neural Network is one of the most utilized system for prediction models, ANN depends on the structure and highlights of Neural Networks, the impersonation of human cerebrum. Right nowcomputational units are called as neurons, these neurons are associated together in layers, where the information is passed in as the information the system is prepared all through with extraordinary

conditions called as the enactment capacities. The utilizations of neural system is broadly utilized for agrarian practices. When the neural system is talented it can expect the harvest yield in equivalent examples, notwithstanding the way that the earlier information comprises of a couple of slip-ups. Regardless of whether the insights are intricate, multivariate, nonlinear this network offers the right impacts and furthermore with no of basic ideas the connection among them and the yield is separated.

VI. Basic Algorithms used

Random Forest Algorithm

Random forest is a sort of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a sort of learning where you join differing kinds of algorithms or same algorithm multiple times to create a more powerful prediction model. The random forest algorithm combines multiple algorithm of a similar type i.e. multiple decision trees, leading to a forest of trees, hence the name "Random Forest".

Support Vector Machine

"Support Vector Machine" (SVM) is a directed AI calculation which can be utilized for both arrangement and relapse difficulties. Be that as it may, it is for the most part utilized in order issues. In the SVM calculation, we plot every datum thing as a point in n-dimensional space (where n is number of highlights you have) with the estimation of each component being the estimation of a specific arrange. At that point, we perform arrangement by finding the hyper-plane that separates the two classes very well. Support Vectors are basically the co-ordinates of individual perception. The SVM classifier is an outskirt which best isolates the two classes (hyper-plane/line).

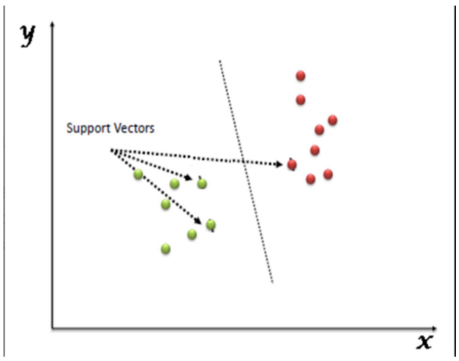


Fig.4: Sample Diagram of Support Vector Machine.

VII. Implementation and Analysis of Results

The proposed SVM based crop yield prediction system is tested for Soybean dataset [XVIII] and the resultant graphs of Least Squares Support Vector Machine predictions are as in Fig. 5.and meteorological parameter predictions as in Fig. 6.

From the analysis, it is concluded that there is a huge yield expectation in rural arrangements as in Fig. 5, while implementing Random Forest Algorithm.

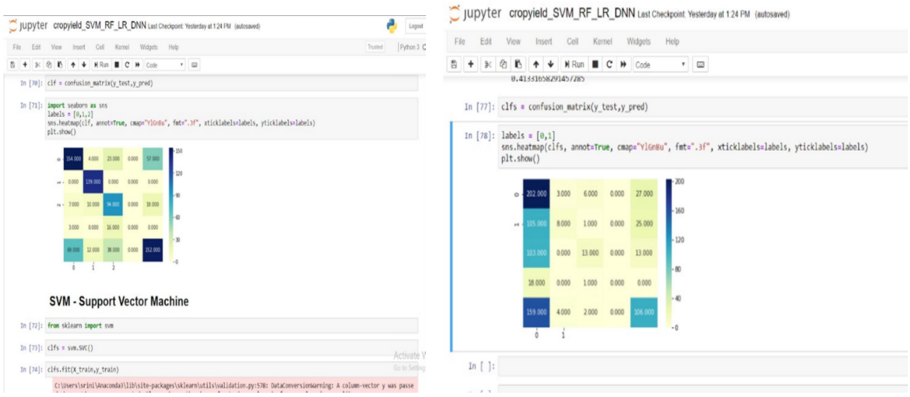
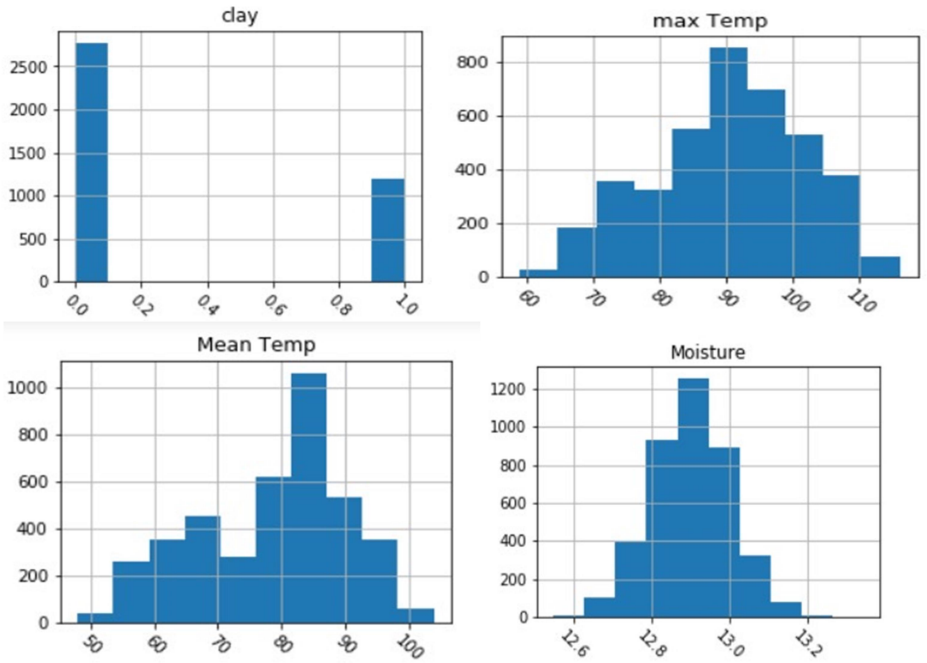


Fig.5: Crop Yield prediction using SVM

As shown in Fig. 6.,parameters of temperature, moisture achieves normal distribution.



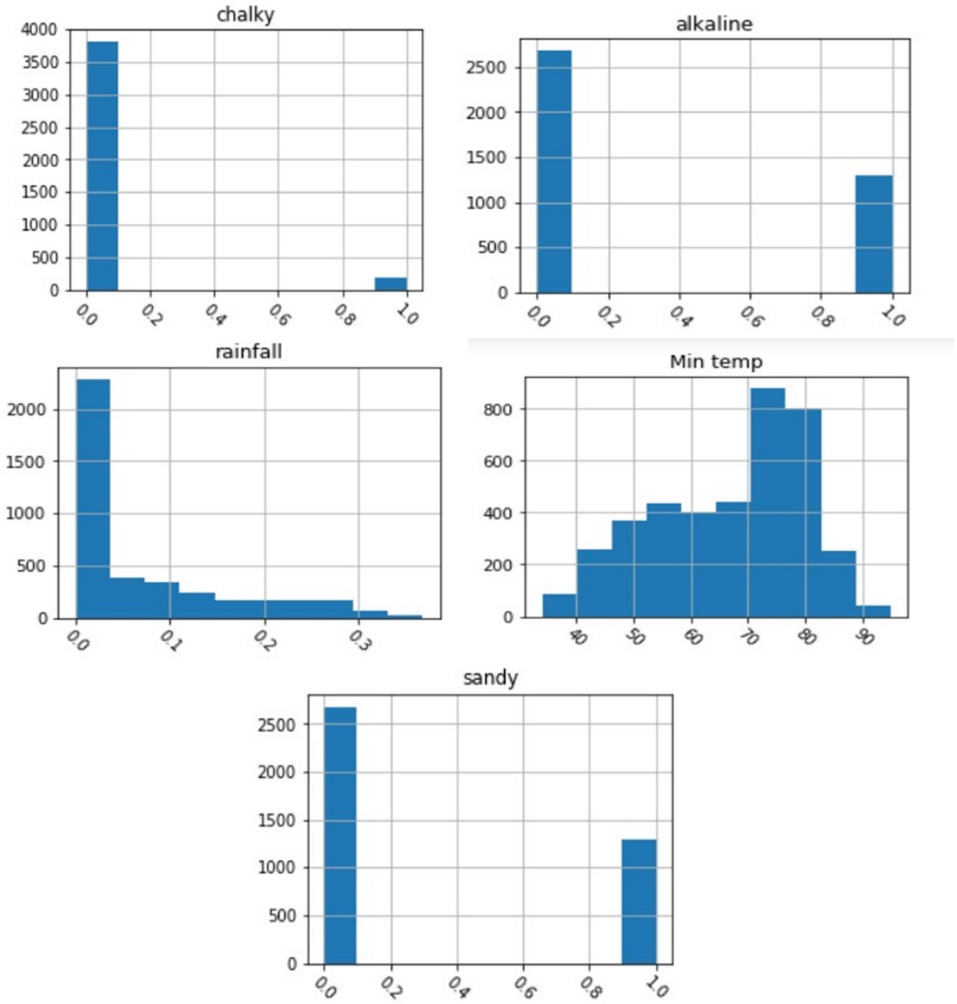


Fig.6: Prediction Graphs on various Meteorological parameters.

VIII. Conclusion and Future Enhancements

The proposed system achieves an improved harvest yield expectation utilizing the Random Forest calculations along with Least Squares Support Vector Machines. Random Forest algorithm accomplishes a biggest number of harvest yield with the most minimal models. It is most appropriate for huge yield expectation in rural arrangements. This makes the ranchers to take the correct choice for right harvest to such an extent that the rural part will be created by inventive thoughts.

Harvest Yield Prediction includes foreseeing yield of the harvest from verifiable and accessible information like climate, soil and other noteworthy harvest yield parameters. To anticipate the harvest yield in future precisely, Random Forest algorithm can still be customized accordingly for ANN and utilized effectively.

References

I. A.Na, W. Isaac, S. Varshney and E. Khan, "An IoT based system for remote monitoring of soil characteristics," 2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds, Noida, 2016, pp.316-320, doi: 10.1109/INCITE.2016.7857638.

II. AnupamaMahato, "Climate Change and its Impact on Agriculture", International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014.

III. Awanit Kumar, Shiv Kumar, "Prediction of production of crops using K-Means and Fuzzy Logic", International Journal of Computer Science and Mobile Computing, Vol.4 Issue.8, August- 2015, pg. 44-56.

IV. Birthal, P.S., Kumar, S., Negi, D.S. and Roy, D. (2015), "The impacts of information on returns from farming: evidence from a nationally representative farm survey in India. Agricultural Economics", 46: 549-561. doi:10.1111/agec.12181

V. Dhivya B, Manjula, Siva Bharathi, Madhumathi, "A Survey on Crop Yield Prediction based on Agricultural Data", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Issue 3, March 2017.

VI. G. Ravichandran and R. S. Koteeshwari, "Agricultural crop predictor and advisor using ANN for smartphones," 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), Pudukkottai, 2016, pp. 1-6. doi: 10.1109/ICETETS.2016.7603053.

VII. https://en.wikipedia.org/wiki/Linear_regression.

VIII. https://en.wikipedia.org/wiki/Nonlinear_regression.

IX. Iooss, Bertrand &Lemaître, Paul.(2014). A Review on Global Sensitivity Analysis Methods.Operations Research/ Computer Science Interfaces Series. 59. 10.1007/978-1-4899-7547-8-5.

X. JapneetKaur," Impact of Climate Change on Agricultural Productivity and Food Security Resulting in Poverty in India", Final Thesis, Master's Degree Programme – Second Cycle, UniversitaCaFoscariVenezia, 2017.

XI. J. Shenoy and Y. Pingle, "IOT in agriculture", 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 1456-1458.

XII. L. Leroux, C. Baron, B. Zoungrana, S. B. Traoré, D. Lo Seen and A. Bégué, "Crop Monitoring Using Vegetation and Thermal Indices for Yield Estimates: Case Study of a Rainfed Cereal in Semi-Arid West Africa," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 1, pp. 347-362, Jan. 2016. doi: 10.1109/JSTARS.2015.2501343.

XIII. M. R. Bendre, R. C. Thool and V. R. Thool, "Big data in precision agriculture: Weather forecasting for future farming," 2015 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, 2015, pp. 744-750. doi: 10.1109/NGCT.2015.7375220.

XIV. M. Paul, S. K. Vishwakarma and A. Verma, "Analysis of Soil Behaviour and Prediction of Crop Yield Using Data Mining Approach," 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, 2015, pp. 766-771. doi: 10.1109/CICN.2015.156.

XV. N. Hemageetha, "A survey on application of data mining techniques to analyze the soil for agricultural purpose," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 3112-3117.

XVI. R. Mythili, MeenakshiKumari, ApoorvTripathi, Neha Pal, "IoT Based Smart Farm Monitoring System", International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume-8 Issue-4, November 2019.

XVII. S. Nagini, T. V. R. Kanth and B. V. Kiranmayee, "Agriculture yield prediction using predictive analytic techniques," 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), Noida, 2016, pp. 783-788. doi: 10.1109/IC3I.2016.7918789.

XVIII. Soybeandataset, "https://archive.ics.uci.edu/ml/datasets.php?format=mat&task=cla&att=&area=life&numAtt=10to100&numIns=&type=mvar&sort=dateUp&view=list".

XIX. Zhihua Zhang, Multivariate Time Series Analysis in Climate and Environmental Research, 2018, Springer Nature Switzerland.