International Journal of Control and Automation
Vol.10, No.7 (2017), pp.13-22
http://dx.doi.org/10.14257/ijca.2017.10.7.02

Adapting Mining into Agriculture Sector with Machine Learning Techniques

A.V.S. Pavan Kumar[1] and R. Bhramaramba[2]

[1]*Research Scholar, Department of Computer Science and Engineering, GIT, GITAM University, Visakhapatnam*
avspavankumarmca@gmail.com
[2]*Associate Professor, Department of Information Technology, GIT, GITAM University, Visakhapatnam*
bhramarambaravi@gmail.com

Abstract

Economy of a country is broadly classified as Industrial and Agricultural economy. An agrarian society like that of India derives the economic strength from its vast agricultural resources. The constant endeavor to improve the effective and efficient yield per hectare with an aim to meet the ever increasing demand for food is a vital factor for our democracy to prosper. Using any and every technique which can optimize agricultural output is a welcome step. Mathematically, the Agricultural yield is formulated by considering host of variables and many models are developed to improve the output. Artificial intelligence methods, such as Machine Learning, are most suitable for development and application of this method in the agricultural scenario. Use of Algorithms and Artificial Intelligence techniques to process the available data is termed as Machine Learning. Machine learning requires human intervention while automated processing of data is being done. In order to accommodate all the conceivable scenarios, Machine Learning adopts a generalized sequence of operations which can be applied to new data. This paper discusses the machine learning techniques like Decision Tree and Support Vector Machines which will be applied on agricultural data. The various techniques demonstrate the benefit to farmers in terms of being cost effective, while strengthening the existing farmers, and adding new farmers to the fold. More importantly, existing models in agriculture can be improved by better correlation of extra variables and developing the non linear relationships.

Keywords: *Machine Learning, Decision Tree, Support Vector Machines, Agriculture Census, Applications of Agriculture.*

1. Introduction

Machine learning methods innovate, great application in Agriculture census [1]. The assorted scope of quickly extending information created by current data science has fuelled a requirement for precise Clustering and forecast calculations. The exactness of characterization calculations are multiple and manifold. The data is expressed as inclusive in certain cases and non inclusive in others depending on the situation under consideration. In machine learning, even a subtle change in atomic level data, would result in a new event. In view of the multitude of parameters to be dealt in the agriculture scenario, this becomes more crucial. The techniques discussed in this section are based on

1. Learning Style.
2. Similarity.

ISSN: 2005-4297 IJCA

1.1. Learning Style

In Learning Style several algorithms can be devised to utilize the data available, this method brings out the difference between natural intelligence and the artificial machine power. In machine learning, the categorization into supervised learning and unsupervised learning, gives us the scope and flexibility of the methods.

1.1.1. Supervised Learning: Supervised Learning can be applied to set up a procedure to perform some forecasts and is also used to overcome when the expected results don't meet the requirements. Here the input is the prepared information and subjected to the algorithm. The calculations are course corrected so as to match the expectations at every stage. For Example:

1. Classification
2. Regression
    (a) Linear Regression
    (b) Logistic regression,
3. Back Propagation Neural Network.

1.1.2. Unsupervised Learning: Unsupervised Learning Mechanism can be used to set up by deriving a proper structure contains in the available information. The algorithm is applied to the model developed based on the perceived structure in the data. A methodical reduction in repetition is applied. For Example:

1. Clustering
2. Hierarchical Clustering
3. Dimensionality reduction, etc.

1.1.3. Semi-Supervised Learning: Semi supervised learning is a blend of the two methods with marked and unlabeled data sets with certain improvisations. Classification and regression models are the best examples. Case calculations are augmentations to other adaptable techniques that consider presumptions on how to present the unlabelled information. While considering this information to model and convert in to business decisions, the most commonly used tools are directed and unsupervised learning techniques. An intriguing issue right now is semi- administered learning techniques in zones, for example, picture arrangement where there are vast datasets with not many named illustrations.

1.2. Similarity

Calculations are frequently gathered by closeness as far as their capacity is concerned. For instance, tree-based strategies, and neural network inspired techniques. This methodology is the most helpful approach to gathering calculations. Though this is a valuable gathering strategy, this cannot be considering the best. There are still calculations that could simply fit into various classes like Learning Vector Quantization that is both a neural system propelled strategy and an occasion based technique. There are additionally classifications that have the same name that depict the issue and the class of calculation, for example, Regression and Clustering. These cases are dealt by posting calculations twice or by selecting the most effective convenient and simple. This last best cluster since there is no need of copying calculations.

Regression is an important technique where in the variables are successively refined by the iteration process so as to narrow down the difference between the estimated and actual values. However, this may lead to confusion as regression is also used to imply the class of issue and the class of estimation. Really, Regression is a methodology. The most surely understood Regression counts are: Ordinary Least Squares Regression (OLSR), Linear Regression ,Logistic Regression, Stepwise Regression, Multivariate Adaptive

Regression Splines (MARS), Locally Estimated Scatter plot Smoothing (LOESS), Occurrence based Algorithms.

Occurrence based learning model, develops a comprehensive data base which utilizes the similarity in existing and new data and locates the best match. A case of Victor takes all strategy with memory based learning. Of the many such examples, KNN, SOM, LVQ, LWL are some of the methods, which have well accepted in the academia.

The variation in method , using the regularization calculations , which are basically adjustments made to different techniques are available in literature like : Ridge Regression, Elastic Net, Least Absolute Shrinkage and Selection Operator (LASSO), Least-Angle Regression (LARS).

Decision tree techniques develop a model of Decisions made taking into account real estimations of qualities in the information. Decisions fork in tree structures until a forecast Decision is made for a given record. Decision trees are prepared on information for Clustering and Regression issues. Decision trees are regularly quick and precise and most loved in machine learning. The most prominent Decision tree calculations are: Classification and Regression Tree (CART), C4.5 and C5.0 (diverse renditions of an intense methodology), Iterative Dichotomiser 3 (ID3), Chi-squared Automatic Interaction Detection (CHAID), M5, Conditional Decision Trees, and Decision Stump.

Bayesian strategies are those that unequivocally apply Bayes Theorem for issues, for example, characterization and Regression. The most well known Bayesian calculations are: Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Averaged One-Dependence Estimators (AODE), Bayesian Belief Network (BBN), and Bayesian Network (BN).

Clustering, similar to Regression depicts the class of issue and the class of strategies. Clustering techniques are commonly sorted out by the displaying methodologies, for example, centroid-based and hierarchical. All techniques are hampered by the use of the innate structures as a part of the information to best sort out the information into clusters of most frequently shared characteristic. The most famous Clustering calculations are: k-Medians, k-Means, Expectation Maximization (EM), Hierarchical Clustering.

Association mining decides learning as strategies that best clarify watched connections between variables in information. These principles can find essential and monetarily helpful relationship in huge multidimensional datasets that can be abused by an association. The most famous Association principles of learning calculations are: Apriori calculation, Eclat calculation.

In the endeavor to develop and mimic natural structures, such as that of the Human Brain, Artificial neural networks were developed. These techniques were used in regression analysis. Excluding the deep learning model, because of its vast expanse, the most researched methods in artificial neural network calculations are based on Perceptron, Hopfield Network, Back-Propagation, and Radial Basis Function Network (RBFN).

Cutting edge techniques of Deep Learning , with their quest to mimic the mind boggling neural networks , the following are some of the prominent methods : : Deep Belief Networks (DBN), Convolutional Neural Network (CNN), Deep Boltzmann Machine (DBM), Stacked Auto-Encoders.

Dimensionality Reduction Algorithms like Clustering strategies, dimensionality reduction look for and abuse the characteristic structure in the information, yet for this situation in an unsupervised way or request to compress or portray information utilizing less data. This can be helpful to envision dimensional information or to improve information which can subsequently be utilized as a part of a supervised learning technique. A large portion of these strategies can be modify an adapted for use in characterization and Regression: Principal Component Regression (PCR), Principal Component Analysis (PCA), Partial Least Squares Regression (PLSR), Linear Discriminant Analysis (LDA), Multidimensional Scaling (MDS), Mixture Quadratic

Discriminant Analysis (QDA), Discriminant Analysis (MDA), and Flexible Discriminant Analysis (FDA).

Aggregation techniques allow us to use weaker models to develop the frame work and then join the same into a more robust larger network having a larger capacity to handle data. Some of the most notable methods are Bootstrapped Aggregation (Bagging), Boosting, AdaBoost, Stacked Generalization (blending), Gradient Boosting Machines (GBM), Gradient Boosted Regression Trees (GBRT) and Random Forest.

2. Literature Survey

Agriculture and the breeding industry are major contributors to our overall economy. It is always a constant endeavor to improve the breeds, improve traits of resistance to parasites and diseases , improve strains which reduce dependence on nutrients , water etc., so that the overall benefit remains positive .

Application of machine learning to agriculture is an incredible advancement in agricultural technologies because it allows these systems to leverage and combine historic information with real- time information such as weather forecasts and vehicle telemetric content, along with the tacit information, the experience that the farmer has. Machine learning techniques augment this heterogeneous data by combining it with the farmer's knowledge to provide coherence to the vast amounts of data being generated in today's farm world.

It allows these systems to learn about characteristics of each field and adapt systemic recommendations that can be improved over a period of time. Machine learning allows farmers and agribusinesses alike to make better decisions, in real-time, even in the absence of complete information.

Ankur M Vyas surveyed distinctive strategies used to recognize fruits based on colour [1]. According to them the most important factor which adopts mechanized fruit grading is the colour of the fruit. The frame work for automated organic product evaluation framework should include a process for shading space and division. This paper dealt in detail the various feature extraction based on colour.

Arivazhagan et al. proposed framework as a programming answer for programmed recognition and order of plant leaf sicknesses [2]. The proposed calculation's effectiveness can effectively recognize and order the inspected infections with a precision of 94%. Around 500 plant leaves were examined and the results conclusively demonstrate the effectiveness of this method.

Rajendra Prasad et al. portray the DM Structure advancement, portrayal, parts utilized for harvest expectation; planting strategist test results are especially helpful to the agriculturists to comprehend advertise needs and planting methodologies [3].

Victor Rodriguez-Galiano et al. evaluated groundwater defenselessness to nitrate contamination utilizing Irregular Forest calculation and indicated technique to include a choice way to deal with decrease in the quantity of explicative factors [4].

Christian Bauckhage, Kristian and Kersting studied late work on computational knowledge in accuracy cultivating [5]. From the perspective of design acknowledgment and information mining, the major challenges in horticultural applications give off an impression of being the accompanying:

1. The across the board arrangement and convenience of present day, (portable) sensor advancements prompts detonating measures of information. This posts the issues of BIG DATA and high throughput calculations which can assimilate Tera bytes of information.
2. Since farming is a genuinely interdisciplinary whose professionals are most certainly not essentially prepared analysts or information researchers, procedures for information investigation should be interpretable and justifiability comes about.

3. Versatile processing for applications “out in the fields” needs to adapt to asset limitations, such as, confined battery life, low computational power, or constrained transmission capacities for information exchange. Calculations expected for portable flag preparing and examination need to address these imperatives. The selected approach in view of a distributional perspective of hyper-otherworldly marks which they utilized for Bayesian expectation of the improvement of dry season pushes levels. They likewise displayed a course of straightforward picture handling and investigation ventures of low computational costs that takes into account dependably recognizing diverse contagious leaf spots in various leaves of beetle plants.

Dr D Ashok Kumar and N Kannathasan have amalgamated the data mining strategies for soil data which results in efficient cultivation, enhanced bio diversity, reduced dependence on manures, superior soil administration.

Farah Khan and Dr. Divakar Singh try to give an outline of some past examines and contemplates done toward applying information mining, particularly, affiliation run mining methods in the agrarian area [7]. In additionally attempted to assess the present status and conceivable future patterns around there. The speculations behind information mining and affiliation principles are introduced first and a study of various procedures connected is given as part of the development. Amina Khatra demonstrated that utilizing shading based picture division yellow rust from wheat can be successfully differentiated. The achievement of the division and genuine entrance of yellow rust for the most part rely on the positioning of the cameras introduced with a specific end goal to procure the pictures from the field. R. D. Tillett in his audit highlighted various ranges of agribusiness area in which picture preparing and distinctive techniques. For example acknowledgment was executed like reaping of fruits like apples, oranges, peaches etc and vegetables like potatoes, carrots, greens etc.

3. Proposed Work

The main objective of this paper is to applying machine learning techniques and there by address the following issues. The steps involved are: Study various algorithms related to Applying Machine Learning on Agriculture data; Study and analyze the various techniques in Machine Learning; Require Adaptation of existing or development of new approaches that give best solution to analysis; Analyze the data using different existing techniques and combine the techniques to get more accurate results; Set up the Data base; Analyze the data to find the relationships in the data; Discovered relationships must be validated; Predictions are used to support decision making process and Come up with best technique for prediction and visualize more accurate results.

The most important actionable points for success are: An increased support for research in public and private sectors, increased investment in agriculture sector, global betterment of agricultural policies, equitable safety net for producers, increased reliance on scientific sanitary rules. Etc.

The main challenges in this work are: Infer knowledge pertaining to Agriculture using Machine Learning Methods, to ascertain their access to leased-in land for crop cultivation, to understand land resources, to predict the cropping pattern and resource utilization. The main steps involved are: Data collection from governmental publications. Lead to inconsistency - need to handle about the accuracy of data, identify the suitable machine learning techniques and apply on processed data. As per the study of the existing papers, Regression techniques are suitable. Based on the data availability and requirements suitable techniques can be listed and if required customized composite methods may be implemented.

Probable solutions to improve are: the proposed work expects the outputs to dovetail into the overall development strategy. It helps to initiate measures to provide greater

access to leased-in land to all kinds of Agriculturists. It provides Knowledgebase to suggest alternative cropping patterns based on available resources. It helps to evolve suitable strategies to improve the existing land resources to identify the relative disadvantages if any in crop production vis-a-vis comparable social groups. 4. Comparative Work the overall machine learning algorithms with its properties summarized are given in table 1. The summary of report is based on data collected.

4. Comparative Work

The overall machine learning algorithms with its properties summarized are given in table 1. The summary of report is based on data collected from [14, 17, 19 and 21].

Table 1. Summary Report on Classification Techniques

Classification Techniques	Attributes	Linear (L)/ Non-Linear (NL)	Feature Ratio /Effect on Sample	Computation on Complexity	Assumptions	Outlier or Noise Effect	Transparency	Incremental Learning
MLP Neural Network	High	NL	Medium	High	None	Low	Poor	Poor
Self - Organizing Maps	Medium	NL	Medium	Medium	None	Low	Poor	Poor
RBF Neural Network	High	NL	Medium	Medium	None	Low	Good	Poor
Probabilistic Neural Network	High	NL	Medium	Medium	None	Low	Good	Poor
Linear Discriminant Analysis	Low	L	Low	Low	Gaussian, equal Variance	Medium	Good	Medium
Support Vector Machines	Low	L/NL	Low	Medium	Variable	Low	Good	Medium
Quadratic Discriminant Analysis	Low	NL	Low	Low	Gaussian, unequal Variance	Medium	Good	Medium
Gaussian Mixture Model	Medium	NL	High	High	Variable	High	Good	Poor
K nearest Neighbor	Low	NL	Low	High	None	Low	Good	Good
Decision Trees	Low	NL	Medium	Medium	None	Low	Good	Poor
Naive Bayes	Low	NL	High	Low	None	Low	Good	Poor
Neuro-Fuzzy Systems	Low	NL	Medium	High	None	Low	Good	Poor
Clustering Tools								
Self – Organizing	Low	-	Medium	High	None	Low	Poor	Medium

Maps								
K-Means	Low	-	High	Medium	Spherical Clusters	High	Poor	Good
Fuzzy c-means	Low	-	High	Medium	Spherical Clusters	High	Poor	Good
Hierarchical Clustering	Medium	-	High	Low	None	Low	Good	Good
Dimensionality Reduction								
Principal Component Analysis	None	L	Low	Low	Gaussian densities	High	Good	Medium
Linear Discriminant Analysis	Low	L	Low	Low	Gaussian densities	High	Good	Poor
Sammon's mapping	Low	NL	Low	High	None	Medium	Poor	Poor
Multi-Dimensional Scaling	Low	NL	Low	Low	None	High	Good	Medium
Independent Component Analysis	Low	NL	Medium	Medium	Variable	High	Good	Medium

Adapting machine learning techniques to agriculture, is a new concept and hence in the nascent stage of development. This is reflected in the trend to delve on the data analysis related issues, need for scientific collection of data and its analysis when machine learning tools are adopted. While many strategies , in isolation can give satisfactory results for the individual problems , such as data splitting, parameter optimization, dealing with missing data, how to train classifiers etc. , research efforts are paving way for a guaranteed statistical validity thereby enhancing quality of research. The classification techniques and their characteristics, with accuracy of parameters, graphically represented are shown in fig 1 [23].

Table 2. Characteristics of Various Classifications

Characteristics	Equal Distance Discretizer	Equal Frequency Discretizer	Class Attribute Contingency Coefficient	Chi2
Supervised / Unsupervised	Unsupervised	Unsupervised	Supervised	Supervised
Top Down / Bottom Up	Top Down	Top Down	Top Down	Bottom Up
Parametric	Yes	Yes	No	Yes
Incremental	No	No	Yes	Yes
Static/ Dynamic	Static	Static	Static	Static

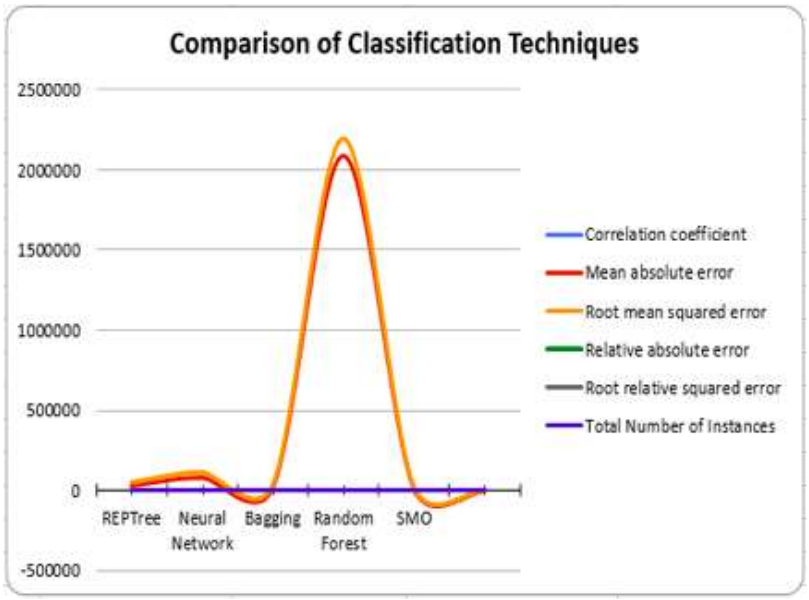

Figure 1. Graphical View of Accuracy Parameters

5. Conclusion

The paper concludes with how machine learning solves agriculture census problems. Agriculture sector shows slightly modest level of adoption of technology and Innovations. Relatively new concepts, which can be implemented with great efficacy even while presenting complex input and output variables. This has been proven successfully with Algorithms already in use in the past in the fields of Computer Science and Statistics. Machine Learning Algorithms had a very positive impact in improving results in Artificial Machines like Sensor based systems applied in precision farming. A review of various applications of the Machine Learning is presented in this study.

References

[1] A. M. Vyas, T. Bjial and N. Sapan. "Colour Feature Extraction Techniques of Fruits: A Survey", International Journal of Computer Applications, vol. 83, no. 15, **(2013)** pp. 15-22.
[2] S. Arivazhagan, R. Newlin Shebiah, S. Ananthi, and S. Vishnu Varthini, "Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features", Agricultural Engineering International, CIGR Journal, vol. 15, no. 1, **(2013)** pp.211-217.
[3] J. R. Prasad, P. R. Prakash, S. S. Kumar, and M. S. B. K. Swarupa Rani, "Identification of Agricultural Production Areas in Andhra Pradesh", International Journal of Engineering and Innovative Technology, (IJEIT), vol. 2, no. 2, **(2012)** pp. 137-140.
[4] V. Rodriguez-Galiano, M. P. Mendes, M. J. Garcia-Soldado, M. Chica-Olmo, and L. Ribeiro, "Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: a case study in an agricultural setting (Southern Spain)", Science of the Total Environment, vol. 476, **(2014)** pp. 189-206.
[5] C. Bauckhage and K. Kersting, "Data mining and pattern recognition in agriculture", KI-Künstliche Intelligenz, Vol. 27, No. 4, pp. 313-324. **(2013)**
[6] A. Kumar, and N. Kannathasan, "A survey on data mining and pattern recognition techniques for soil data mining", IJCSI International Journal of Computer Science Issues, vol. 8, no. 3, **(2011)** pp. 422-428.
[7] F. Khan and Divakar Singh, "Association rule mining in the field of agriculture: a survey", International Journal of Scientific and Research Publications, vol. 329, **(2014)**.
[8] A. Khatra, "Yellow Rust Extraction in Wheat Crop based on Color Segmentation Techniques", IOSR Journal of Engineering (IOSRJEN), vol. 3, no. 12, **(2013)** pp. 56-58.
[9] R. D. Tillett, "Image analysis for agricultural processes: a review of potential opportunities", Journal of agricultural Engineering research, vol. 50, **(1991)** pp. 247-258.

[10] A. Chinchuluun, P. Xanthopoulos, V. Tomaino, and P. M. Pardalos, "Data mining techniques in agricultural and environmental sciences", New Technologies for Constructing Complex Agricultural and Environmental Systems, IGI Global, **(2012)** pp. 311-325.
[11] N. Bhatt and P. V. Virparia, "A Survey based Research for Data Mining Techniques to Forecast Water Demand in Irrigation", International Journal of Computer Science and Mobile Applications, Vol.3, No.8, **(2015)** pp. 14-18.
[12] D. D. Gutierrez, "Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R", Technics Publications, (**2015**).
[13] R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh., "Crop Selection Method to maximize crop yield rate using machine learning technique", 2015 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), **(2015)** pp. 138-145.
[14] M. P. Raj, P. R. Swaminarayan, J. R. Saini, and D. K. Parmar, "Applications of Pattern Recognition Algorithms in Agriculture: A Review", International Journal of Advanced Networking and Applications, vol. 6, no. 5, **(2015)** pp. 2495-2502.
[15] L. K. Mehra, C. Cowger, K. Gross, and P. S. Ojiambo, "Predicting pre-planting risk of Stagonospora nodorum blotch in winter wheat using machine learning models", Frontiers in plant science, Vol. 7 , Article 390, **(2016)** pp. 1-14.
[16] P. Mehta, H. Shah, V. Kori, V. Vikani, S. Shukla, and M. Shenoy., "Survey of unsupervised machine learning algorithms on precision agricultural data", 2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), **(2015)** pp. 1-8.
[17] S. S. Dahikar and S. V. Rode., "Agricultural crop yield prediction using artificial neural network approach", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, vol. 2, no. 1, **(2014)** pp. 683-686.
[18] C. S. Nandyala and H.-K. Kim, "Big and Meta Data Management for U-Agriculture Mobile Services", International Journal of Software Engineering and Its Applications, Vol. 10, No. 2 , **(2016)** pp. 257-270.
[19] T. O. Ayodele, "Types of machine learning algorithms", INTECH Open Access Publisher, (**2010**).
[20] D. Ramesh and B. V. Vardhan, "Data mining techniques and applications to agricultural yield data", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 9, **(2013)** pp. 3477-80.
[21] S. Ghosh and S. Koley, "Machine Learning for Soil Fertility and Plant Nutrient Management using Back Propagation Neural Networks", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 2, no. 2, pp.292-297.
[22] C. Szepesvári,, "Algorithms for reinforcement learning", Synthesis lectures on artificial intelligence and machine learning, vol. 4, no. 1, **(2010)** pp. 1-103.
[23] A. Savla, H. Bhadada, P. Dhawan, and V. Joshi., "Application of machine learning techniques for yield prediction on delineated zones in precision agriculture", International Journal of New Computer Architectures and Their Applications, vol. 5, no. 2, **(2015)** pp. 48-53.
[24] W. Okori and J. Obua., "Machine learning classification technique for famine prediction", Proceedings of the world congress on engineering, vol. 2, **(2011)** pp. 991-996.
[25] http://farmer.gov.in/mspdet.html.

