ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/355026706

An Investigation on Crop Yield Prediction Using Machine Learning

Conference Paper · September 2021

DOI: 10.1109/ICIRCA51532.2021.9544815

CITATIONS
79

READS
1,058

6 authors, including:

Guna Sekhar Sajja
University of the Cumberlands
53 PUBLICATIONS **1,572** CITATIONS
SEE PROFILE

Subhesh Saurabh Jha
Banaras Hindu University
15 PUBLICATIONS **307** CITATIONS
SEE PROFILE

Mohd Naved
Jaipuria Institute of Management
117 PUBLICATIONS **3,091** CITATIONS
SEE PROFILE

Samrat Ray
Peter the Great St.Petersburg Polytechnic University
426 PUBLICATIONS **4,467** CITATIONS
SEE PROFILE

All content following this page was uploaded by Samrat Ray on 05 June 2022.

The user has requested enhancement of the downloaded file.

Proceedings of the Third International Conference on Inventive Research in Computing Applications (ICIRCA-2021)
IEEE Xplore Part Number: CFP21N67-ART; ISBN: 978-0-7381-4627-0

AN INVESTIGATION ON CROP YIELD PREDICTION USING MACHINE LEARNING

Guna Sekhar Sajja[1]
Research Scholar
University of the Cumberlands
guna.sajja@gmail.com

Subhesh Saurabh Jha[2*]
Research Scholar, Department of Botany, Institute of Sciences, Banaras Hindu University
subheshs.jha2@bhu.ac.in

Hicham Mhamdi[3]
Laboratory of Electronic Systems, Information Processing, Mechanics and Energetics, Faculty of Sciences, Kenitra, University Ibn Tofail Kenitra, Kenitra, Morocco
hicham.mhamdi@uit.ac.ma

Dr. Mohd Naved[4]
Assistant Professor,
Department of Business Analytics, Jagannath University, Delhi-NCR, India.
mohdnaved@gmail.com

Samrat Ray[5]
Research scholar, The Institute of Industrial Management, Economics and Trade, Peter The Great Saint Petersburg Polytechnic University,Russia.
samratray@rocketmail.com

Khongdet Phasinam[6]
School of Agricultural and Food Engineering, Faculty of Food and Agricultural Technology, Pibulsongkram Rajabhat University, Phitsanulok, Thailand
phasinam@psru.ac.th

ABSTRACT:

For the existence of humans, agriculture is vitally crucial. For a big population of the globe, agriculture provides a living. It also provides the locals with a large number of work openings. Many farmers desire to use old-fashioned farming techniques, which provide poor income. Critical to the economy's long-term development and advancement are agriculture and the related industries. Decision making, crop selection and supporting systems for increased crop output are the primary problems for agricultural production. The prediction of agriculture depends on parameters such as temperature, soil fertility, amount of water, water quality and seasons, crop price, etc. Machine learning plays an important role in crop yield prediction on the basis of geography, climate details, and season. It helps farmers in growing most appropriate crop for their farm land. This paper presents a machine learning based framework for prediction of crop yield. For experimental set up, a data set is created for crop details. Machine learning algorithms SVM, random forest and ID3 are used for investigation.

Keywords- Machine learning, Crop Yield Prediction, SVM, ID3, Random Forest, Classification, Prediction, Preprocessing, Ensemble Learning

I. INTRODUCTION

One of the tactics proposed by the UN Council to achieve this aim is to increase the worldwide supply of high-quality food. However, the unrestrained population growth necessitates novel solutions to the challenge. Prediction of human population and crop output is one approach to dealing with the problem.

Accurate crop yield forecasting during the growing season provides various benefits to policymakers and farmers, including projecting market pricing, planning import and export, and reducing the social effect of crop loss. Timely food production choices improve national food security. Agricultural entrepreneurs and smallholders profit from such forecasts as well, because they can make educated decisions about managing and financing their crops [1]. Because of the intricacy of the data, crop production forecast is a difficult challenge for policymakers. Agriculture and agroeconomics academics are very interested in creating new mathematical approaches that provide better prediction utilizing the existing parameters. The study in this field is concerned with presenting a link between the area of agriculture and crop output, while taking into account many environmental variables such as soil quality, irrigation, and how land is managed. These models are based on rules with parameters. The professionals engaged have a keen understanding of the types of correlations that may be identified with the factors involved in agriculture and the environment.

The difficulty arises when it is realized that such knowledge cannot be defined for the construction of empirical expert system rules. Crop production [2] is predicted using manual surveys and remote sensing data. Manual studies using historical knowledge of prior years' observations with mathematical tools are beneficial for a localized area, but they are difficult to extend to other areas and nations. Recent advancements in crop simulation models have solved these issues. Crop simulation models use mathematical models of soil conditions, weather,

2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) | 978-1-6654-3877-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIRCA51532.2021.9544815

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on June 05,2022 at 14:39:37 UTC from IEEE Xplore. Restrictions apply.

and management strategies to mimic crop development throughout the growing season. To predict agricultural production [3] across wide regions, these simulation methods require a huge dataset. Data is frequently obtained utilizing remote sensing technologies such as satellites, aircraft, drones, or a basic camera.

The construction of empirical models to predict agricultural production has long been acknowledged as an essential task for the remote sensing community. Traditional (linear) statistical modeling proved difficult due to the systems' complexity and non-linearity [4]. The research community discovered in the 1990s that nonlinear models may provide a more realistic and presumably more accurate answer to the difficult issue of empirical yield modeling. Artificial Neural Networks [5] and decision trees [6] first appeared in agricultural yield prediction using empirical modeling in the early 2000s. According to the most recent literature, ANNs continue to be popular for empirical agricultural yield prediction. More contemporary machine learning (ML) approaches [7] [8], such as Support Vector Machines (SVM) and Random Forest (RF), have piqued the interest of the scientific community and are being developed for use in data science.

II. RELATED WORK

Using soil quality metrics and a tillage system, MLR and ANN algorithms were used to estimate organic potato production. The impact of tillage practices on soil parameters while calculating crop yield is explored (Abrougui [9]). They discovered that tillage and soil conditions had a significant influence on yield. The crop production was likewise calculated more precisely by the MLR model than by the ANN model. Nonetheless, its prediction accuracy was lower than that of the ANN model.

Bocco et al. [10] estimated daily global solar radiation in a portion of Argentina's Salta Province using linear and statistical models. MLR, ANN, and Multilayer Perceptron were used to assess the dataset's features. The linear models and neural network models were created, and their efficacy was evaluated using the dataset. They utilized a data collection that contained information regarding solar radiation statistics from 1996 to 2002. Three different combinations of meteorological parameters were investigated for neural networks and linear regressions. Both prediction approaches yielded positive findings for the researchers.

Sahin et al. investigated the estimate capabilities of MLR and ANN [11]. Using algorithms, the researchers estimate Turkey's average solar radiation in this study. The satellite was used to collect data from 73 different places. The algorithms were used to examine data gathered from satellites and meteorological sources. The monthly average radiation, GPS coordinates, and land surface temperature were used as input characteristics in ANN and MLR to predict average solar radiation. The findings demonstrated that the ANN model outperformed the MLR model in terms of performance.

Mohammad Zaefizadeh et al. [12] evaluated MLR and ANN to estimate barley output. Their prediction model consisted of 15 neurons and was based on multilayer ANN with one hidden layer. The Matlab Perceptron software utilized in this work was based on an algorithm that employed an error propagation learning mechanism and a hyperbolic tangent function. The investigation' comparing results revealed that the mean deviation index of estimate in the ANN methodology was one-third of its MLR rate. The variance in mean deviation index value was caused by a substantial interaction between genotype and environment. This interaction has an effect on the MLR estimate approach. This study indicated that a neural network technique was preferable to regression for yield prediction, especially when there were substantial genotype-environment interactions and higher velocity.

Safa and Samarasinghe [13] aimed to develop an ANN capable of predicting energy use in wheat production. During the 2007-08 harvest season, the study was conducted on both irrigated and dry wheat fields in Canterbury. Extensive interviews and questionnaires were used to obtain data. Many direct and indirect elements were found by the researchers in order to train the ANN. When a dataset was chosen for testing and validation, the ANN model predicted energy use better than the MLR model.

Gonzalez S anchez, Frausto Sol s, Ojeda Bustamante et al. [14] investigated the predictive performance of ML and linear regression techniques in crop yield prediction utilizing data obtained from a Mexican irrigation zone utilizing 10 crop datasets. In addition to the MLR model, the researchers examined the prediction abilities using regression trees, neural networks, closest neighbor, and support vector models. M5-Prime achieved the greatest average accuracy matrices and k-nearest neighbor approaches, and the study concluded that in agricultural planning, the planner might utilize the tool M5-prime to forecast higher crop output.

O. Satir et al. [15] developed an estimates display for edit development that makes use of the vegetation purpose of files and the Stepwise Linear Regression (SLR) show. Furthermore, the region's related trim assortments were con-organized using a multitransient Landsat data collection and object-based categorization. In this case, an ongoing metric such as Mean Percent Error (MPE) forecasts the yields. The MPE forecasted crops such as corn, cotton, and wheat, which were computed and used using an uneven measure of salinity in the soil. This investigation, the forecast, was completed employing only a climate data, as a result, taking only a single attribute.

Pritam et al. [16] developed a spiking model based on Neural Network (NN) for the purpose of calculating crop production and Spatio-temporal analysis of time series images. The system is made up of highly parallel hardware platforms with low power consumption neuromorphic properties. The SNN

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on June 05,2022 at 14:39:37 UTC from IEEE Xplore. Restrictions apply.

computational model is inferred for estimating trim yield from standardized distinction vegetation list image time arrangement. The testing of a methodical system has been completed, as well as the spatial collection of time arrangement from a Moderate Resolution Imaging Spectro-radiometer 250-m determination information and conventional crop yield information to construct an SNN to provide convenient harvest yield prediction. To maximize the findings from the experimental data set, studies on the optimal amount of characteristics are also supplied. The approach estimates the yield with excellent precision about a month and a half before collection. In light of a nine feature model, our methodology offered a normal exactness of 95.64 percent, with a normal blunder of expectation of 0.236 t/ha and a relationship coefficient of 0.801.

Campos et al. [17] calculated agricultural water productivity and yield as a simplified remote sensing-driven technique of approaching the real process using satellite data. It makes use of the key that covers biophysical parameters. According to the FAO-66 Aquacrop Handbook, the link between edit transpiration and biomass generation is ensured. The findings of the researcher's examination show the association between biomass output and transpiration coefficient. Similarly, multiple studies have demonstrated a greater association between Kcb and remote sensing-based VI. As a result, the association between the two factors is well established. Using recorded data collected over an 11-year period, the approach assesses the connection between biomass production and reflectance as a function of Kcb. The study verifies the existence of a significant association and lays the path for the use of remote sensing data in a quantitative analysis of agricultural biomass production and yield.

E.I. Papageorgiou et al. [18] investigated the yield expectation technique in cotton edit generation using the delicate registration approach of fluffy intellectual maps. Fuzzy Cognitive Map was created by combining fuzzy techniques with subjective map ideas. This was used to demonstrate and speak to the learning of specialists. It was capable of coping with circumstances with unclear descriptions by employing a technique comparable to that of human thinking. It was a difficult decision-making methodology, especially in complicated processing systems. In light of the application concept, the FCM methodology demonstrated here was employed in horticulture. Cotton production was a complex method with several cooperating factors, and FCMs were appropriate for this type of problem. FCM was written and designed to communicate with master data for cotton crop forecasting and harvest management. The investigated concept was evaluated for 360 cases measured over a six-year period (2001-2006) on a 5 ha test cotton field, in forecasting the yield class between two possible categories ("low" and "high"). The evaluation results revealed its identical favorable position over the benchmarking machine learning calculations attempted for comparable informational index for the years stated by offering decisions that match better with the genuine measured ones.

Luke Bornn and James V. Zidek [19] demonstrated how spatial dependency might be merged into quantitative models for trim yield while avoiding the drawbacks of ignoring it. A Bayesian framework for trim yield was created using completely stamped biophysical logical items and it employs spatially-settled prior learning of probability conveyances. It considered extended exhibiting flexibility and also for improved want over existing minimum squares procedures. The approach was designed to provide profitable estimates that account for the effects of wild diversions. The previous circulations were constructed to accommodate the spatial non-attractive expanding from diverse district disparities in agricultural arrangement. As a result, the model developed superior prediction execution close to standard models and enabled coordinated knowledge of climatic influences on the model's yield.

III. METHODOLOGY FOR CROP YIELD PREDICTION

A. FRAMEWORK

As shown in figure 1, a framework for crop yield prediction is presented. This framework consists of a crop yield data set. The data set is preprocessed using data cleaning techniques. Then a set of machine learning techniques is applied on the data set to perform classification. This help in prediction of crop type and duration for a particular field.

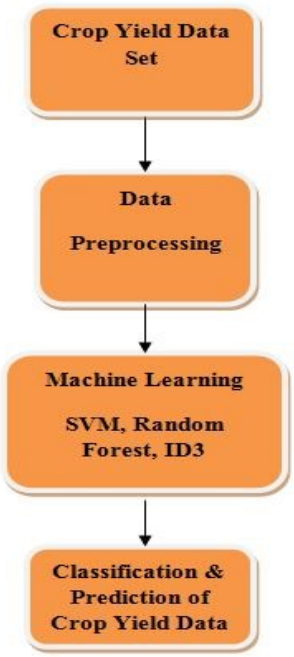

Fig.1 Crop Yield Prediction Using Machine Learning

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on June 05,2022 at 14:39:37 UTC from IEEE Xplore. Restrictions apply.

The Support Vector Machine is a non-probabilistic binary linear classification approach. It builds a training model that divides the data into one or more target classes. The data objects are represented as points in space. A noticeable gap separates the items of distinct categories, causing its width to spread. The new instances' target classes are mapped based on which side of the gap they arrive on. Non-linear classification is also possible with the support vector machine when the input datasets are not labeled. Because there are no target classes to which the instances may be mapped, the support vector machine uses an unsupervised learning methodology for categorizing data Following the formation of clusters based on functions, new instances are added to them. The author [20] describes an effective model-based recommendation system based on non-linear support vector machine. Non-linear support vector machine approaches are the most widely utilized methodology for dealing with unlabeled data and are utilized in a wide range of industrial applications.

A researcher named J. Ross Quinlan created the ID-3 technique (Iterative Dichotomiser-3), which is the first developing decision tree-based approach. This method is based on measures of entropy and information gain. The original dataset begins with a base nodule and computes the entropy of the functional features for each iteration. The attribute with the lowest error rate (entropy) and the greatest information gain is chosen as a split attribute, and the dataset is divided to produce a subset of attributes based on it. Unless the procedure is correctly classified to its target classes, it is repeated recursively on every subset of data. The decision tree is constructed using a nonterminal node, and the terminal nodes are specified by the ultimate subset of the branch. The split attribute defines the nonterminal node, whereas the terminal node represents the class labels. To efficiently identify and anticipate cardiac issues at an early stage, it leverages an ID-3-based decision tree algorithm created by [21].

Author [22] introduced Random Forest as an ensemble learning strategy for classification and regression operations in his article. During the training phase, it generates a huge number of decision trees and using regression techniques to predict the outcomes of the individual trees. It has a low variance and connects the different features of the presented data fast for prediction purposes. The reason for the initial lack of enthusiasm for this methodology is because random forest categorization approaches are difficult to grasp.

B. RESULT ANALYSIS

A crop details data set of 750 instances is prepared. This data set consists of attributes like year, region name, crop (cotton, groundnut, jowar, rice and wheat.), season (kharif, rabi, summer), area (in hectares), production (in tonnes), average temperature (°C), average rainfall (mm), soil, PH value, soil type, major fertilizers, nitrogen (kg/Ha), phosphorus (Kg/Ha),Potassium(Kg/Ha), minimum rainfall required, minimum temperature required. class (Predicts the type of crop and duration of crop). This data preprocessing has helped in improving accuracy. The accuracy and error rate achieved is shown below in figure 2 and figure 3.

Accuracy= (TP + TN) / (TP + TN + FP + FN)

Where

TP= True Positive

TN= True Negative

FP= False Positive

FN= False Negative

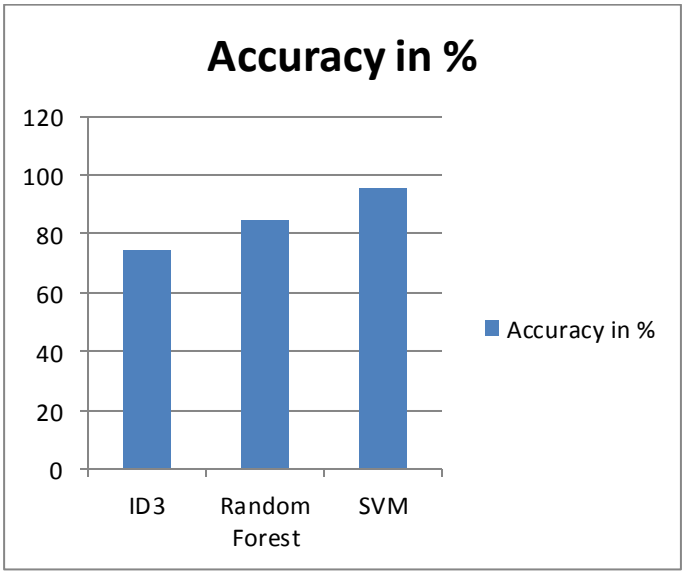

Fig.2 Accuracy Results of Classification Algorithms

Error rate is calculated using following formulae.

Error Rate = (FP + FN)/ (TP + TN + FP + FN)

Where

TP= True Positive

TN= True Negative

FP= False Positive

FN= False Negative

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on June 05,2022 at 14:39:37 UTC from IEEE Xplore. Restrictions apply.

Fig.3 Error Rate Results of Classification Algorithm

IV. Conclusion

Agriculture is critical to the survival of humanity. Agriculture provides a living for a large portion of the world's population. It also gives a great number of job opportunities for locals. Many farmers want to continue using old-fashioned farming practices that generate a low revenue. Agriculture and allied businesses are critical to the economy's long-term development and prosperity. The key difficulties for agricultural production are decision making, crop selection, and supporting systems for enhanced crop output. Agriculture forecasting is influenced by variables such as temperature, soil fertility, water availability, water quality and season, crop price, and so on. Machine learning is useful in predicting agricultural yields based on location, climatic data, and season. It assists farmers in cultivating the best suitable crop for their agricultural property. This research describes a machine learning-based system for agricultural yield prediction. A data collection for crop details is prepared for experimental setup. SVM, random forest, and ID3 machine learning techniques are employed for inquiry. SVM has outperformed other machine learning algorithms in accuracy

References

[1] A. Nigam, S. Garg, A. Agrawal and P. Agrawal, "Crop Yield Prediction Using Machine Learning Algorithms," 2019 Fifth International Conference on Image Information Processing (ICIIP), 2019, pp. 125-130, doi: 10.1109/ICIIP47207.2019.8985951.

[2] F. F. Haque, A. Abdelgawad, V. P. Yanambaka and K. Yelamarthi, "Crop Yield Prediction Using Deep Neural Network," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1-4, doi: 10.1109/WF-IoT48130.2020.9221298.

[3] Srinivasulu, A., Ramanjaneyulu, K., Neelaveni, R. et al. Advanced lung cancer prediction based on blockchain material using extended CNN. Appl Nanosci (2021).

[4] S. Kothapalli, M. Samson, S. Majji, T. R. Patnala, S. R. Karanam and C. S. Pasumarthi, "Comparative Experimental Analysis of different Op-amps using 180nm CMOS Technology," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-6, doi: 10.1109/ic-ETITE47903.2020.440.

[5] Tulasi Radhika Patnala, Jayanthi D, Shylu D.S, Kavitha K, Prathyusha Chowdary, "Maximal length test pattern generation for the cryptography applications" https://www.sciencedirect.com/science/article/pii/S2214785320305368, meterialstoday proceedings, In press, available online from 20.02.2020

[6] Tulasi Radhika Patnala, Jayanthi D, Sankararao Majji, Manohar Valleti, Srilekha Kothapalli, Santhosh Chandra Rao Karanam, "Modernistic way for KEY Generation for Highly Secure Data Transfer in ASIC Design Flow" https://ieeexplore.ieee.org/document/9074200, Published in IEEE digital Xplore, Electronic ISSN: 2575-7288, available from 23.04.2020

[7] R. Medar, V. S. Rajpurohit and S. Shweta, "Crop Yield Prediction using Machine Learning Techniques," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033611.

[8] Haoxiang, Wang, and S. Smys. "Big Data Analysis and Perturbation using Data Mining Algorithm." Journal of Soft Computing Paradigm (JSCP) 3, no. 01 (2021): 19-28

[9] M. Sivakami, P. Prabhu, Classification of Algorithms Supported Factual Knowledge Recovery from Cardiac Data Set, International Journal of Current Research and Review, vol.13. issue 6. pp161-166. ISSN: 2231-2196 (Print) ISSN: 0975-5241 (Online).

[10] Dr Ashim Bora. Dr. N.Vasantha Gowri.,Dr. Mohd Naved.,Dr.Purnendu Shekhar Pandey. (2021). An Utilization Of Robot For Irrigation Using Artificial Intelligence. International Journal of Future Generation Communication and Networking, 14(1).

[11] Prabhu, P.,Selvabharathi, S. (2019), Deep Belief Neural Network Model for Prediction of Diabetes Mellitus. In 2019 3rd International Conference on Imaging, Signal Processing and Communication, ICISPC 2019 (pp. 138–142) Institute of Electrical and Electronics Engineers Inc. ISBN:9781728136639.

[12] Ganesh, R. S., Jausmin, K. J., Srilatha, J., Indumathy, R., Naved, M., & Ashok, M. (2021, April). Artificial Intelligence Based Smart Facial Expression Recognition Remote Control System. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1056-1061). IEEE.

[13] Abrougui, K., Gabsi, K., Mercatoris, B., Khemis, C., Amami, R. and Chehaibi, S. (2019), 'Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ann) and multiple linear regressions (mlr)', Soil and Tillage Research 190, 202–208

[14] Bocco, M., Willington, E., Arias, M. et al. (2010), 'Comparison of regression and neural networks models to estimate solar radiation', Chilean Journal of Agricultural Research 70(3), 428–435.

[15] S. ahin, M., Kaya, Y. and Uyar, M. (2013), 'Comparison of ann and mlr models for estimating solar radiation in turkeyusing noaa/avhrr data. advances in space research 51 (2013) 891-904'.

[16] Fortin, J. G., Anctil, F., Parent, L.-E. and Bolinder, M. A. (2011), 'Site-specific early ´season potato yield forecast by neural network in eastern canada', Precision agriculture 12(6), 905–923.

[17] Safa, M. and Samarasinghe, S. (2011), 'Determination and modelling of energy consumption in wheat production using neural networks:"a case study in canterbury province, new zealand"', Energy 36(8), 5140–5147.

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on June 05,2022 at 14:39:37 UTC from IEEE Xplore. Restrictions apply.

[18] Gonzalez-Sanchez, A., Frausto-Solis, J. and Ojeda-Bustamante, W. (2014), 'Attribute selection impact on linear and nonlinear regression models for crop yield prediction', The Scientific World Journal 2014
[19] O. Satir, and S. Berberoglu, "Crop yield prediction under soil salinity using satellite derived vegetation indices," Field Crops Research, vol.192, pp.134-143, 2016.
[20] P. Bose, N.K. Kasabov, L. Bruzzone, and R.N. Hartono, "Spiking Neural Networks for Crop Yield Estimation Based on Spatiotemporal Analysis of Image Time Series," IEEE Transactions on Geoscience and Remote Sensing, vol.54, no.11, pp.6563-6573, 2016.
[21] I. Campos, C.M. Neale, T.J. Arkebauer, A.E. Suyker, and I.Z. Gonçalves, "Water productivity and crop yield: A simplified remote sensing driven operational approach," Agricultural and Forest Meteorology, 2017.
[22] E.I. Papageorgiou, A.T. Markinos and T.A. Gemtos, "Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application", vol.11, no.4, pp.3643-3657, 2011.
[23] L. Bornn, and J.V. Zidek, "Efficient stabilization of crop yield prediction in the Canadian Prairies", International journal of agricultural and forest meteorology, vol.152, pp.223-232, 2012.
[24] Zhang, Shuai, Y.-L. S. A. (2017), 'Deep learning based recommender system: a survey and new perspectives', Journal of ACM Computing Surveys 1(1), 1–35.
[25] Hssina, Merbouha, A. and Ezzikouri (2014), 'A comparative study of decision tree ID3 and C4.5', International Journal of Advanced Computer Science and Applications 4(2), 13–19.
[26] Cheng-Hsiung Wenga, Tony Cheng-Kui Huang, R.-P. H. (2016), 'Disease prediction with different types of neural network classifiers', Journal of Telematics and Informatics (4), 277–292.

Authorized licensed use limited to: Peter the Great St. Petersburg Polytechnic Univ. Downloaded on June 05,2022 at 14:39:37 UTC from IEEE Xplore. Restrictions apply.
View publication stats