IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.3, Mar 2020 E-ISSN: 2347-2693

A Machine Learning Approach to Predict Crop yeild and Reduction of Cost by Finding Best Accuracy

Spoorthi P[1*], Jayashankara M[2]

[12] Department of Computer Science & Engineering, PES College of Engineering, Mandya, Karnataka, India

DOI: https://doi.org/10.26438/ijcse/v8i3.6166 | Available online at: www.ijcseonline.org

Received: 27/Feb/2020, Accepted: 15/Mar/2020, Published: 30/Mar/2020

Abstract— Among worldwide, agriculture has the major responsibility for improving the economic contribution of the nation. However, still the most agricultural fields are under developed due to the lack of deployment of ecosystem control technologies. Due to these problems, the crop production is not improved which affects the agriculture economy. Hence a development of agricultural productivity is enhanced based on the plant yield prediction. To prevent this problem, Agricultural sectors have to predict the crop from given dataset using machine learning techniques. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the best crop. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with entropy calculation, precision, Recall, F1 Score, Sensitivity, Specificity.

Keywords—*Dataset,Machine learning-classification method*

I. INTRODUCTION

In developing countries, farming is considered as the major source of revenue for many people. In modern years, the agricultural growth is engaged by several innovations, environments, techniques and civilizations. In addition, the utilization of information technology may change the condition of decision making and thus farmers may yield the best way. For decision making process, data mining techniques related to the agriculture are used. Data mining is a process of extracting the most significant and useful information from the huge amount of datasets. Nowadays, we used machine learning approach with developed in crop or plant yield prediction since agriculture has different data like soil data, crop data, and weather data. Plant growth prediction is proposed for monitoring the plant yield effectively through the machine learning techniques.

It is also applicable for the automated process of farming is the beginning of a new era in Bangladesh that will be suitable for the farmers who seek experts to take suggestion about the appropriate crop on specific location of their land and don't want to forget any step of the cultivation throughout the process. Although, the opinion from experts is the most convenient way, this application is designed to give accurate solution in fastest manner possible. This research's main objective is to bring farming process a step closer to the digital platform.

We have used Machine learning to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data.

II. RELATED WORK

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

[1]Becker-Reshef , E. Vermote, M. Lindeman , C. Justice Wheat is one of the key cereal crops grown worldwide, providing the primary caloric and nutritional sourcefor millions of people around the world. In order to ensure food security and sound, actionable mitigation strategies and policies for management of food shortages, timely and accurate estimates of global crop production are essential. This study combines a new BRDF-corrected, daily surface reflectance dataset developed from NASA's Moderate resolution Imaging Spectro-radiometer (MODIS) with detailed official crop statistics to develop an empirical, generalized approach to forecast wheat yields. The first step of this study was to develop and evaluate a regression-based model for forecasting winter wheat production in Kansas. This regression-based model was then directly applied to forecast winter wheat production in Ukraine. The forecasts of production in Kansas closely matched the

USDA/NASS reported numbers with a 7% error. The same regression model forecast winter wheat production in Ukraine within 10% of the official reported production numbers six weeks prior to harvest. Using new data from MODIS, this method is simple, has limited data requirements, and can provide an indication of winter wheat production shortfalls and surplus prior to harvest in regions where minimal ground data is available.

[2]M.S. Mkhabelaa, P. Bullocka, S. Rajb, S. Wangc, Y. Yangc

In this paper the Normalised Difference Vegetation Index (NDVI) data derived from the advanced very high resolution radiometer (AVHRR) sensor have been extensively used to assess crop condition and yield on the Canadian Prairies and elsewhere, NDVI data derived from the new moderate resolution imaging spectroradiometer (MODIS) sensor have so far not been used for crop yield prediction on the Canadian Prairies. Therefore, the objective of this study was to evaluate the possibility of using MODIS-NDVI to forecast crop yield on the Canadian Prairies and also to identify the best time for making a reliable crop yield forecast. Growing season (May–August) MODIS 10-day composite NDVI data for the years 2000–2006 were obtained from the Canada Centre for Remote Sensing (CCRS). Crop yield data (i.e., barley, canola, field peas and spring wheat) for each Census Agricultural Region (CAR) were obtained from Statistics Canada. Correlation and regression analyses were performed using 10-day composite NDVI and running average NDVI for 2, 3 and 4 dekads with the highest correlation coefficients (r) as the independent variables and crop grain yield as the dependent variable. To test the robustness and the ability of the generated regression models to forecast crops grain yield, one year at a time was removed and new regression models were developed, which were then used to predict the grain yield for the missing year. Results showed that MODIS-NDVI data can be used effectively to predict crop yield on the Canadian Prairies. Depending on the agro-climatic zone, the power function models developed for each crop accounted for 48 to 90%, 32 to 82%, 53 to 89% and 47 to 80% of the grain yield variability for barley, canola, field peas and spring wheat, respectively, with the best prediction in the semi-arid zone. Overall (54 out of 84), the % difference of the predicted from the actual grain yield was within ±10%. On the whole, RMSE values ranged from 150 to 654, 108 to 475, 204 to 677 and 104 to 714 kg ha^{-1} for barley, canola, field peas and spring wheat, respectively. When expressed as percentages of actual yield, the RMSE values ranged from 8 to 25% for barley, 10 to 58% for canola, 10 to 38% for field peas and 6 to 34% for spring wheat. The MAE values followed a similar trend but were slightly lower than the RMSE values. For all the crops, the best time for making grain yield predictions was found to be from the third dekad of June through the third dekad of July in the sub-humid zone and from the first dekad of July through the first dekad of August in both the semi-arid and arid zones. This means that accurate crop grain yield forecasts using the developed regression models can be made one to two months before harvest.

[3]Douglas K. Bolton, Mark A. Friedl

In this paper we used data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) in association with county-level data from the United States Department of Agriculture (USDA) to develop empirical models predicting maize and soybean yield in the Central United States. As part of our analysis we also tested the ability of MODIS to capture inter-annual variability in yields. Our results show that the MODIS two-band Enhanced Vegetation Index (EVI2) provides a better basis for predicting maize yields relative to the widely used Normalized Difference Vegetation Index (NDVI). Inclusion of information related to crop phenology derived from MODIS significantly improved model performance within and across years. Surprisingly, using moderate spatial resolution data from the MODIS Land Cover Type product to identify agricultural areas did not degrade model results relative to using higher-spatial resolution crop-type maps developed by the USDA. Correlations between vegetation indices and yield were highest 65–75 days after greenup for maize and 80 days after greenup for soybeans. EVI2 was the best index for predicting maize yield in non-semi-arid counties (R2 = 0.67), but the Normalized Difference Water Index (NDWI) performed better in semi-arid counties (R2 = 0.69), probably because the NDWI is sensitive to irrigation in semi-arid areas with low-density agriculture. NDVI and EVI2 performed equally well predicting soybean yield (R2 = 0.69 and 0.70, respectively). In addition, EVI2 was best able to capture large negative anomalies in maize yield in 2005 (R2 = 0.73). Overall, our results show that using crop phenology and a combination of EVI2 and NDWI have significant benefit for remote sensing-based maize and soybean yield models.

[4]Sabareeswaran and R. Gunasundari

Among worldwide, agriculture has the major responsibility for improving the economic contribution of the nation. However, still the most agricultural fields are under developed due to the lack of deployment of ecosystem control technologies. Due to such issue 6, the crop production is not improved which affects the agriculture economy. Hence in this paper, a development of agricultural productivity is enhanced based on the plant yield prediction. Initially, different features such as plant images, soil characteristics, and weather factors are gathered and Firefly (FF) optimization algorithm is proposed for Feature Selection (FFFS). Then, the most selected optimal features are classified based on the Modified Fuzzy Cognitive Map (MFCM) algorithm for predicting the growth of plant yield. The predicted outcome is transmitted to the farmer's through smart phones which helps for identifying the growth of plant and improving the harvesting. The experimental results show that the effectiveness of the proposed technique can be compared with the other prediction techniques.

[5]Paul C. Doraiswamy, Sophie Moulin, Paul W. Cook, and Alan Stern.

Monitoring crop condition and production estimates at the state and county level is of great interest to the U.S. Department of Agriculture. The National Agricultural Statistical Service (NASS) of the U.S. Department of Agriculture conducts field interviews with sampled farm operators and obtains crop cuttings to make crop yield estimates at regional and state levels. NASS needs supplemental spatial data that provides timely information on crop condition and potential yields. In this research, the crop model EPIC (Erosion Productivity Impact Calculator) was adapted for simulations at regional scales. Satellite remotely sensed data provide a real-time assessment of the magnitude and variation of crop condition parameters, and this study investigates the use of these parameters as an input to a crop growth model. This investigation was conducted in the semi-arid region of North Dakota in the southeastern part of the state. The primary objective was to evaluate a method of integrating parameters retrieved from satellite imagery in a crop growth model to simulate spring wheat yields at the sub-county and county levels. The input parameters derived from remotely sensed data provided spatial integrity, as well as a real-time calibration of model simulated parameters during the season, to ensure that the modeled and observed conditions agree. A radiative transfer model, SAIL (Scattered by Arbitrary Inclined Leaves), provided the link between the satellite data and crop model. The model parameters were simulated in a geographic information system grid, which was the platform for aggregating yields at local and regional scales. A model calibration was performed to initialize the model parameters. This calibration was performed using Landsat data over three southeast counties in North Dakota. The model was then used to simulate crop yields for the state of North Dakota with inputs derived from NOAA AVHRR data. The calibration and the state level simulations are compared with spring wheat yields reported by NASS objective yield surveys.

The scope of this project is to investigate a dataset of crop records for agricultural sector using machine learning technique. To identifying crop predicting by farmer is more difficult. We try to reduce this risk factor behind selection of the crop.

Data collection:

The data set collected for predicting past farmer list of yield is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Random Forest, logistic, Decision tree algorithms are applied on the Training set and based on the test result accuracy, Test set prediction is done.

Data cleaning/preparing process:

Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

III. METHODOLOGY

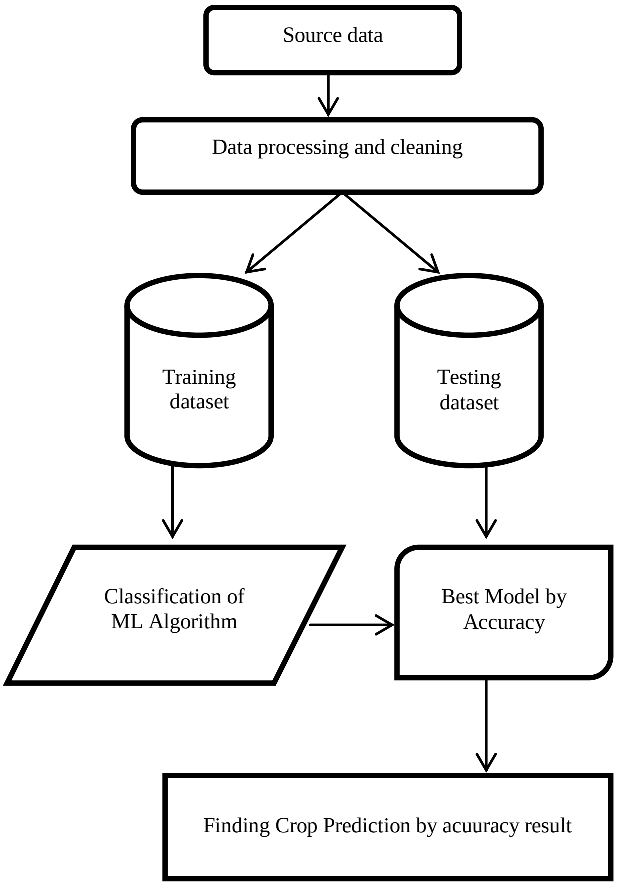
Source data
Data processing and cleaning
Training dataset
Testing dataset
Classification of ML Algorithm
Best Model by Accuracy
Finding Crop Prediction by acuuracy result

Figure 1.Flow chart of the process

Training the Dataset:

- The first line imports iris data set which is already predefined in sklearn module. Iris data set is basically a table which contains information about various varieties of iris flowers.
- For example, to import any algorithm and train_test_split class from sklearn and numpy module for use in this program.
- Then we encapsulate load_data() method in data_dataset variable. Further we divide the dataset into training data and test data using train_test_split method. The X prefix in variable denotes the feature values and y prefix denotes target values.

- This method divides dataset into training and test data randomly in ratio of 67:33. Then we encapsulate any algorithm.
- In the next line, we fit our training data into this algorithm so that computer can get trained using this data. Now the training part is complete.

Testing the Dataset:

- Now we have dimensions of a new flower in a numpy array called 'n' and we want to predict the species of this flower. We do this using the predict method which takes this array as input and spits out predicted target value as output.
- So the predicted target value comes out to be 0. Finally we find the test score which is the ratio of no. of predictions found correct and total predictions made. We do this using the score method which basically compares the actual values of the test set with the predicted values.

This helps all others department to carried out other formalities. It have to find Accuracy of the training dataset, Accuracy of the testing dataset, Specification, False Positive rate, precision and recall by comparing algorithm using python code.

Comparing Machine learning algorithms:
Before that comparing algorithm, Building a Machine Learning Model using install Scikit-Learn libraries. In this library package have to done preprocessing, linear model with logistic regression method, cross validating by KFold method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the train set and test set. To predicting the result by comparing accuracy.

Prediction result by accuracy:
Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression or other model by comparing the best accuracy.

IV. RESULTS

The final outcome gives the best accuracy, precision, recall and F1 score,sensitivity,specificity.

Prediction result by accuracy:

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.
True Positive Rate(TPR) = TP / (TP + FN)
False Positive rate(FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

(a)Accuracy = (TP + TN) / (TP + TN + FP + FN)
Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

(b)Precision: The proportion of positive predictions that are actually correct.
Precision = TP / (TP + FP)
Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

(c)Recall: The proportion of positive observed values correctly predicted.
Recall = TP / (TP + FN)
Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

(d)F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula:
F- Measure = 2TP / (2TP + FP + FN)
F1-Score Formula:
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

(e)Sensitivity: Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive). Sensitivity is also termed as Recall. This implies that there will be another proportion of actual positive cases, which would get predicted incorrectly as negative (and, thus, could also be termed as the false negative).

Mathematically, sensitivity can be calculated as the following:

Sensitivity = (True Positive) / (True Positive + False Negative)

The following is the details in relation to True Positive and False Negative used in the above equation.

- True Positive =The true positive represents the number of persons who are unhealthy and are predicted as unhealthy.

- False Negative = The false negative represents the number of persons who are unhealthy and got predicted as healthy. Ideally, we would seek the model to have low false negatives as it might prove to be life-threatening or business threatening.

(f)Specificity: Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative). This implies that there will be another proportion of actual negative, which got predicted as positive and could be termed as false positives.

Mathematically, specificity can be calculated as the following:

Specificity = (True Negative) / (True Negative + False Positive)

The following is the details in relation to True Negative and False Positive used in the above equation.

- True Negative =The true negative represents the number of persons who are healthy and are predicted as healthy.

- False Positive =The false positive represents the number of persons who are healthy and got predicted as unhealthy.

The below table shows the performance of Machine learning parameters,the end results gives us the accuracy of crop yield prediction and crop yield cost prediction.

Table 1: Comparison of accuracy results of crop yield prediction:

Parameters	LR	DT	RF	SVC
Precision	0.96	1	1	0.62
Recall	0.96	1	1	1
F1-Score	0.96	1	1	0.77
Sensitivity	0.95	1	1	1
Specificity	0.92	1	1	0
Accuracy (%)	94.73	100	100	62.06

Table 2: Comparison of accuracy results of crop yield cost prediction:

Parameters	LR	DT	RF	SVC
Precision	0.70	1	1	0.69
Recall	0.88	1	1	1
F1-Score	0.78	1	1	0.82
Sensitivity	0.87	1	1	1
Specificity	0.13	1	1	0
Accuracy (%)	65.17	100	100	69.36

Advantages:

Our goal is push for assisting farmers, government using our predictions. All these publications state they have done better than their competitors but there is no article or public mention of their work being used practically to assist the farmers. If there are some genuine problems in rolling out that work to next stage, then identify those problems and try solving them. It is targeted to those farmers who wish to professionally manage their farm by planning, monitoring and analyzing all farming activities.

IV. CONCLUSION AND FUTUREWORK

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finally we predict the crop using machine learning algorithm with different results. The best accuracy on public test set is higher accuracy score by Machine learning method from calculating cross validation checking, Precision, recall and F1score in future. This brings some of the following insights about crop prediction. As maximum types of crops will be covered under this system, farmer may get to know about the crop which may never have been cultivated and lists out all possible crops, it helps the farmer in decision making of which crop to cultivate. Also, this system takes into consideration the past production of data which will help the farmer get insight into the demand and the cost of various crops in market.

In the future, Remaining SMLT algorithms will be involve to finding the best accuracy with applying to predict the crop yield and cost. Agricultural department wants to automate the detecting the yield crops from eligibility process (real time).To automate this process by show the prediction result in web application or desktop application.To optimize the work to implement in Artificial Intelligence environment.

IV. REFERENCE

[1]Robotic Agriculture, ser. UK-RAS White Papers. UK-RAS Network,2018.

[2]Priyanka R.R., Mahesh M., Pallavi., Jayapala G., Pooja M.R.Crop protection by an alert based system using deep learning concept Research Paper|Isroset(IJSRCSE)Vol.6,Issue.6, pp47-49,Dec-2018

[3] P. Lottes, M. Hoeferlin, S. Sander, M. Muter, P. Schulze, and L. C. ¨Stachniss, "An effective classification system for separating sugar beetsand weeds for precision farming applications," in ICRA, 2016, pp. 5157–5163.

[4] A. Milioto, P. Lottes, and C. Stachniss, "Real-time semantic segmentation of crop and weed for precision agriculture robots leveragingbackground knowledge in cnns," in ICRA, 2018, pp. 2229–2235.

[5] P. Lottes, J. Behley, A. Milioto, and C. Stachniss,"Fully convolutionalnetworks with sequential information for robust crop and weeddetectionn precision farming," IEEE Robotics and Automation Letters (RA-L),vol. 3, pp. 3097–3104, 2018.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net:Convolutional networksfor biomedical image segmentation," in International Conference onMedical image computing and computer-assisted intervention, 2015, pp.234–241.

[7] K.Simonyan and A.Zisserman,"Very deep convolutional networks forlarge-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[8] F.-M. De Rainville, A. Durand, F.-A. Fortin, K. Tanguy, X. Maldague,B. Panneton, and M.-J. Simard, "Bayesian classification and unsupervised learning for isolating weeds in row crops," Pattern Analysis andApplications, vol. 17, no. 2, pp. 401–414, 2014.

[9] S. Haug, A. Michaels, P. Biber, and J. Ostermann, "Plant classificationsystem for crop/weed discrimination without segmentation," in WACV.IEEE, 2014, pp. 1142–1149.

[10] P. Lottes, R. Khanna, J. Pfeifer, R. Siegwart, and C. Stachniss, "Uavbased crop and weed classification for smart farming," in ICRA, 2017,pp. 3024–3031.

[11] S. H. Lee, C. S. Chan, P. Wilkin, and P. Remagnino, "Deep-plant: Plantidentification with convolutional neural networks," in ICIP. IEEE, 2015