Adjustment Pattern of pH Using Random Forest Regressor for Crop Modelling of NFT Hydroponic Lettuce

M N Mara*, Helmy, A Nursyahid, T A Setyawan, A Sriyanto

Electrical Engineering Department, Politeknik Negeri Semarang, Semarang, 50246, Indonesia

*E-mail: muhlasah@polines.ac.id

Abstract. IoT makes it possible to automatically adjust pH based on sensor numerical data for NFT hydroponic lettuce. In this research, we use sensors of light intensity, farm temperature, farm humid, reservoir level, reservoir temperature, humid on the outside farm, outside farm temperature at Kartika Farm and Turus Asri. Data were taken for five cropping periods. We got different adjustment pattern of pH based on these sensors and Total Dissolved Solids (TDS) control. Furthermore, Kartika Farm’s cropping was different from Turus Asri. In a kilogram of Kartika Farm cropping ingredient seven holes of hydroponic lettuces while Turus Asri cropping ingredient ten holes of hydroponic lettuces. Economically, Kartika Farm harvests more benefits than Turus Asri. In this research, we modeling the pattern of pH adjustment at Kartika Farm using random forest regression. The result showed that the important variable for pH adjustment modeling is water level, TDS and green house humidity. We got an accurate value for the model is 78.37% with MAE and MSE consecutively 0.19 and 0.15.

1. Introduction

Hydroponics is a modern technology in farming that uses water as a growing medium. NFT (Nutrient Film Technique) is a type of hydroponic cultivation with plant roots in the shallow water layer. The water contains nutrients and is circulated continuously using a pump, while the roots that are not submerged take oxygen for plant growth. Thus, plants always get adequate nutrition and oxygen needs. The NFT technique enables high quality agricultural products with a shorter time period than other techniques [1].

In hydroponic cultivation, the nutrients used are in the form of a solution. Plant nutrition is very important in the growth process so that the plant needs such as macro and micro nutrients can be fulfilled properly. Therefore, nutrition must be in the right composition. The success of the hydroponic culture system depends on the nutrition provided so as not to cause excessive uptake [2].

A lot of research is done to obtain the right nutrient measure. [3] conducted an F test to analyze the impact of nutrient density combined with water flow. The result of the study was that there was no effect on the interaction between nutrient sensitivity and water flow. [4] conducted research to find out the needs of TDS mustard pakcoy based on the day with fuzzy logic control method. The results of their study are mustard pak coy age 10 days requires EC 3.5 mS/cm and TDS 1750 ppm. [5] used ANOVA to analyze the impact of pH modification on plant growth. The pH modifier used is
lowering the pH, lemon juice and vinegar. The results showed that lemon water can be used to raise the pH in slada production. [6] used random forest for analysing water resource availability for wildlife in forest. [7] showed that Models based on RF algorithms proved to be more accurate for normally distributed. [8] used Random Forest Algorithm to detect the pH level. [9] used Random Forest for helping farmer for managing natural resource. In their research, random forest algorithm gave RMSE value 0.07. This result helps farmers in adopting best management practices and trying to give them the economical and technical support in making easier for them to adopt best management practices.

The data in this study was taken from Kartika Farm and Turus Asri. Both farms are located in Semarang. Kartika farm harvest is different from Turus Asri. Kartika Farm harvests 7 holes/kg while Turus Asri harvests 10 holes/kg. In this study, data visualization was conducted with the plotnine library on pythons to analyze the behavior of TDS, pH and light intensity of kartika farm and Turus Asri crops. In addition, modeling of crops using random forest regression. The dependent variable on that model is pH. While the independent variables are EC, TDS, light intensity, temperature farm, humi farm, reservoir sensor, reservoir level, reservoir temperature, humi outside farm, temperature outside farm.

2. Methods

2.1. Data Collection

There are 10 data variables taken from Kartika Farm and Turus Asri namely pH, TDS, light intensity, farm temperature, farm humidity, sensor reservoir, level reservoir, temperature reservoir, outside farm humidity, outside farm temperature. Data is collected using sensors. the sensors used in this study were GY-302, SHT-21 sensor, HCSR-04pH sensor, EC sensor, and DS18B20 sensor. NodeMCU is used to send all sensor measurement data to the server database over the internet, be it sensors mounted on nodeMCU boards or sensors mounted on nano arduino boards. NodeMCU microcontroller consists of hardware in the form of ESP8266 E-12 module that has wifi feature, allowing connection to the internet. The GY-302 sensor is a light intensity sensor module that uses the BH1750FVI chip with an I2C series connection. This sensor has a digital output signal with a lux output unit. The maximum measurement limit of light intesnistas by this sensor is 65,535 lux. The DS18B20 sensor is a waterproof temperature gauge sensor that can be used to measure temperature in wet places.

2.2. Data Visualization

The early stages of this research are feature engineering. All data is visualized with the python's plotnine library. This is done to analyze pattern data from both farms. In addition, there is also a correlation between variables from both Kartika Farm and Turus Asri. This stage is required as a predictive power of data. The next stage is to use a random forest regressor to predict the pH threshold and the TDS threshold. Random forest regressor is only charged on Kartika Farm data. Random forest (RF) is a algorithm used in the classification of large amounts of data. The classification of random forest is done through tree merging by training on the sample data owned.

2.3. Random Forest Regression

Leo Breiman proposed a shceme call random forest in 2001 [10]. He proposed random forest for building for building a predictor ensemble with a set of decision trees. The decision tree grows in randomly selected subspaces of data. Regression forests are for nonlinear multiple regression. The importance of the predictor variables can be viewed from regression forest. Definition of random forest is a classifier consisting of a collection of trees structured classifiers \( \{ h(x, \Theta_k), k=1, ... \} \) where the \( \{ \Theta_k \} \) are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.
Random forest regresses output is a numeric value instead of a class label as in the random forest classifier. The training set is taken independently of random vector \( y, X \). Mean square error from predictor \( h(x) \) is

\[
E_{XY} = (Y - h(X))^2
\]  

Random forest predictors are obtained by averaging the \( k \) tree \{ \( h(x, \Theta_k) \) \}.

\[
\hat{h}(x) = \frac{1}{k} \sum_{j=1}^{k} \hat{h}_j(x)
\]  

With \( \hat{h}_j(x) \) is a tree estimator based on subsample or a size bootstrap using randomly selected \( p \) features. Generalization of random forest regressor errors can be calculated with the following formula

\[
PE^*_{\text{forest}} = E_{XY}(Y - \Theta h(X, \Theta))^2
\]  

The defined average of error generalizations is

\[
PE^*_{\text{tree}} = E_{\Theta}[E_{XY}(Y - h(X, \Theta))^2]
\]  

Bagging is used for random feature selection. Each training set is taken with a replacement from the original training set. Then a tree is planted on a training set using a selection of random features. There are two reasons for using bagging, namely the first to use bagging to improve accuracy when random features are used. The second bagging is used to provide an estimate of the generalization error \( (PE^*) \) of the combined tree, to estimate strength and correlation. The simplest Random Forest with random features is formed with random selection, on each node, a small group of shared input variables.

3. Result and Discussion

In this section will be outlined the results of the research that has been done. Visualization is performed to see the pattern of setting the \( pH \) and TDS against the light intensity. It appears that there is a different pattern between Kartika farm data and Turus Asri. At Kartika farm, TDS conditions at \( pH \) between 5.5-7.5 are more stable which is about 1500 ppm for light intensity of 0 to 5000 lux. While in Turus Asri for the same level of light intensity, TDS looks stable around 0 to 500 ppm at \( pH \) between 7.5-8.5.

![Figure 1. pH and TDS Pattern Against Light Intensity](image)

Given that the crop from Kartika farm is more profitable in terms of economy, then next will be made modeling the \( pH \) pattern of Kartika Farm using random forest regressor. Therefore in this study the \( pH \) became variable dependent and 9 other variables became independent variables. The data is divided into two parts namely 70% for training data and the remaining 30% as data testing. Obtained accuracy for
this model is 82.7% with MAE values of 0.15 and MSE 0.11. Next, it will be checked whether all independent variables are important for modeling pH. This importance value can be seen in Table 1 below.

### Table 1. Importance Variable

| Variable              | Gini Importance Value |
|-----------------------|-----------------------|
| Green House Humidity  | 0.225541              |
| TDS                   | 0.225292              |
| Reservoir Level       | 0.189196              |
| Light Intensity       | 0.144938              |
| Greenhouse Temperature| 0.080491              |
| Reservoir Temperature | 0.042724              |
| Reservoir Sensor      | 0.022041              |
| Outside Temperature   | 0.000000              |
| Outside Humidity      | 0.000000              |

On Table 1 it is noticeable that the outer temperature and outer humidity are not importance so that it can be removed from the model. Random forest regressor models without variable out temp and out humidity provide 82.25% accuracy. It does not differ significantly from the previous model. When taken only the 5 most important variables namely humidity farm, TDS, reservoir level, light intensity and farm temperature are obtained model accuracy value of 81.1% with MAE 0.16 and MSE 0.12. Accuracy differs by about 1.15% from the model with the 8 independent variables we obtained earlier. Overall the accuracy value for the 4 most important variables up to 2 most important variables can be seen in Table 2. It appears that the accuracy of models with 3 independent variables is not much different from the accuracy of models with more independent variables. However, the accuracy of models with 2 independent variables is quite significant with the accuracy of the model with 3 most important independent variables. Is also for MAE and MSE values. Therefore, the model is taken by entering only 3 most important variables, namely Greenhouse humidity, TDS, reservoir level that measure water level.

### Table 2. Models Accuracy

| Variable Independence                                      | Accuracy | MAE  | MSE  |
|------------------------------------------------------------|----------|------|------|
| Greenhouse humidity, TDS, level reservoir, light, greenhouse temperature sensor reservoir, temperature reservoir, outside greenhouse humidity, outside greenhouse temperature | 82.7%    | 0.149| 0.105|
| Greenhouse humidity, TDS, level reservoir, light, greenhouse temperature sensor reservoir, temperature reservoir | 82.25%   | 0.153| 0.109|
| Greenhouse humidity, TDS, level reservoir, light, greenhouse temperature | 81.1%    | 0.166| 0.124|
| Greenhouse humidity, TDS, level reservoir | 79.6%    | 0.181| 0.142|
| Greenhouse humidity, TDS, level reservoir | 78.37%   | 0.187| 0.146|
| Greenhouse humidity, TDS | 62.04%   | 0.330| 0.271|

Because the regression tree is quite large, then in Figure 2 only the part is displayed for illustration. From the picture it appears that if the water level is 15 cm, TDS 850 ppm and humidity green house 85, then the pH should be set to be worth 6.
Figure 2. Random Forest Regression Decision Tree

4. Conclusion
Importance variables included in the model as variable independent are water level, TDS and level reservoir. Inserting more than three importance variables into the model does not provide a significant accuracy value compared to simply inserting three importance variables into the model. The model obtained had 78.37% accuracy with MAE and MSE consecutively 0.19 and 0.15.

References
[1] M J S M Lenord and C Sridevi 2014 Design of efficient hydroponic nutrient solution control system using soft computing based solution grading Computation of Power, Energy, Information and Communication (ICCPEIC)
[2] Velázquez L A, Hernández M A, León M, Domínguez R B, and Gutiérrez J M 2013 First advances on the development of a hydroponic system for cherry tomato culture 10th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)
[3] H J Celso, R Roberto, S L F Paulo, C A G Antonio, and A F Jose 2008 Effect of electric conductivity, ionic concentration and flow of nutrient solutions in the production of hydroponic lettuce Ciência agrotecnologia 32(1)
[4] Mashumah S, Rivai M, and Irfansyah A N 2019 Nutrient film technique based hydroponic system using fuzzy logic control International Seminar on Intelligent Technology and Its Applications (ISITIA)
[5] Singh H, Dunn B, and Payton M E 2019 Hydroponic pH Modifiers affect Plant Growth and Nutrient Content in Leafy Greens Journal of Horticultural Research 27(1)
[6] Seenivasan M and Chakravarthi V J 2019 Analysis of Water Resource Availability for Wildlife in Forest Using Advanced Random Forest Algorithm AIP Conference Proceedings 2177, 020086
[7] Shichkin A V, Buevich A G, and Sergeev A P 2018 Comparison of artificial neural network, random forest and random perceptron forest for forecasting the spatial impurity distribution AIP Conference Proceedings 1982, 020005
[8] Pavithra K, Jayalakshmi M 2020 Analysis of Precision Agriculture based on Random Forest Algorithm by using Sensor Networks International Conference on Inventive Computation Technologies
[9] Leo Breiman 2001 Random Forest Machine Learning