Erciyes Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 37, Sayı 3 ,2021

ErciyesUniversity
Journal of Institue Of Scienceand Technology
Volume 37, Issue 3, 2021

Comparison of Machine Learning Regression Models for the Prediction of Soil Moisture with the use of Internet of Things Irrigation System Data

Bilal Babayigit[1], **Belkıs Buyukpatpat**[*2],

[1] Erciyes Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği, KAYSERİ
[*2] Bartın Üniversitesi Mühendislik, Mimarlık ve Tasarım Fakültesi Bilgisayar Mühendisliği, BARTIN

(Alınış / Received: 07.06.2021, Kabul / Accepted: 30.12.2021, Online Yayınlanma / Published Online: 30.12.2021)

Keywords
Soil Moisture Prediction,
Machine Learning Regression Models,
Internet of Things,
ThingSpeak

Abstract: Internet of Things (IoT) technology allows the control and management of systems independent of humans. IoT-based agriculture applications have become widespread as a solution to the problems of food consumption and water scarcity in agriculture as the world population has increased gradually. Soil moisture is an important factor for agriculture production and hydrological cycles and the prediction of soil moisture is required in developing agricultural practices. In this study, an IoT-based irrigation system prototype is presented which consists of Esp8266 Wifi module, humidity and temperature, soil moisture, rain and ultraviolet sensors connected to the Arduino Uno board. Using the prototype system, data are collected from the pilot area determined in half-hour periods for 55 days and saved the cloud with ThingSpeak. The soil moisture value is estimated by applying different machine learning regression models such as multiple linear, polynomial, support vector, decision tree and random forest regression using the collected data. To examine the success of the algorithms, the obtained results are compared according to the coefficient of determination and the mean square error criteria. It is found that the random forest regression model has found to be superior to other machine learning algorithms for soil moisture estimation.

Nesnelerin İnterneti Sulama Sistemi Verileri Kullanılarak Toprak Neminin Tahmini için Makine Öğrenimi Regresyon Modellerinin Karşılaştırılması

Anahtar Kelimeler
Toprak Nem Tahmini,
Makine Öğrenmesi Regresyon Modelleri,
Nesnelerin İnterneti,
ThingSpeak

Öz: Nesnelerin İnterneti (IoT) teknolojisi, sistemlerin insanlardan bağımsız olarak kontrol edilmesine ve yönetilmesine olanak tanır. IoT tabanlı tarım uygulamaları, dünya nüfusunun giderek artmasıyla tarımda gıda tüketimi ve su kıtlığı sorunlarına çözüm olarak yaygınlaşmıştır. Toprak nemi, tarımsal üretim ve hidrolojik döngüler için önemli bir faktördür ve tarımsal uygulamaların geliştirilmesinde toprak neminin tahmin edilmesi gerekmektedir. Bu çalışmada Arduino Uno kartına bağlı Esp8266 Wifi modülü, nem ve sıcaklık, toprak nemi, yağmur ve ultraviyole sensörlerinden oluşan IoT tabanlı bir sulama sistemi prototipi sunulmuştur. Sunulan prototip sistemi kullanılarak pilot bir alandan 55 gün boyunca yarım saatlik periyotlarla veriler toplanmış ve bu veriler ThingSpeak ile bulut üzerinden kaydedilmiştir. Toplanan veriler kullanılarak çoklu doğrusal, polinomal, destek vektörü, karar ağacı ve rastgele orman regresyonu gibi farklı makine öğrenimi regresyon modelleri uygulanarak toprak nem değeri tahmin edilmiştir. Bu algoritmaların başarımını incelemek için elde edilen sonuçlar belirlilik katsayısı ve ortalama kare hatası kriterlerine göre karşılaştırılır. Rastgele orman regresyon modeli toprak nem tahmini için diğer makine öğrenmesi algoritmalarından daha üstün bulunmuştur.

***İlgili yazar, e-mail**: bbuyukpatpat@bartin.edu.tr

1. Introduction

With the rapid developments in network technology, the number of objects connected to the Internet is increasing day by day. In direct proportion to this increase, applications and application areas based on the IoT, which can carry out things independently of people by facilitating people's lives, have developed. With the development of application areas, the volume of data produced by IoT applications has also increases [1]. IoT applications are network technology in which interconnected devices that can collect data for continuous control of systems, communicate with each other, can be remotely monitored or controlled. IoT technology enables the world to be smart and appears in many areas such as smart city, smart environment, smart grids, security and emergencies, smart retail, home automation, animal husbandry and smart agriculture [2].

One of the most basic human needs is water and water scarcity has become a universal problem due to the increasing world population [3]. This problem requires the optimum use of available water resources. Food demand dominates the water consumption as it uses 85% of agricultural water resources worldwide. [4]. In addition to preventing the consumption of resources, smart agriculture has an important role in contributing to environmental sustainability and reducing costs. Systems that irrigate periodically used in agricultural activities neglect external factors that may affect soil moisture. Such a system causes the consumption of water resources as well as the inefficiency of the soil by over irrigation. It is important to accurately measure the relationship between soil and water in smart farming practices. The parameter that directly reflects the water requirement of the soil in agricultural applications is soil moisture. For this reason, accurate soil moisture estimation can give an idea about the irrigation needs and the suitability of the area for agriculture in the long term. In the field of smart agriculture, which is one of the IoT application areas, many studies have been carried out to remotely monitor environmental data, provide an information system for agriculture, and collect basic data for smart precision irrigation [5-9].

Liu [10] developed an IoT-based land monitoring system that includes sensors for CO_2, temperature and humidity, soil moisture, light intensity, and pH value. He was used a camera to record photos and videos of his field of study. The data received from the sensors with the microcontroller were transmitted to the web portal and the data was stored in the SQL database ZigBee technology was used for communication, solar energy and battery power were used for the devices used.

Yang et al. [11] proposed an IoT-based greenhouse system where environmental data from sensors such as humidity, temperature, amount of light, CO_2, O_2, 03, NO_2 were stored and analyzed with Hadoop. In the proposed greenhouse system, data sent with ESP8266 Wifi module was stored in MySQL and transferred to the cloud platform and analyzed with Hadoop HDFS. The analyzed data was presented to the user via the web and android interface.

Shekhar et al. [12] proposed a smart irrigation system to determine the water demand of the area using the K-nearest neighbor algorithm (KNN). Humidity and temperature data received from the environment were sent to Raspberry Pi via the gateway. The collected data and the predicted data with the KNN algorithm were stored and displayed on the cloud server.

Gorthi and Dou [13] made a comparison of different models developed for the prediction of soil moisture content using the data obtained from soil moisture experiments between June 23 and July 12. Soil moisture content was used as a model output. The models developed include basic regression models such as polynomial regression, adaptive basic function generation, Bayesian frame-based relation vector machines, multivariate adaptive regression based on iterative partitioning, classification and regression trees based on continuous class learning, and multi-layer perceptron networks. the neural network based model was more efficient than other models with 0.0024 mean square error (MSE) value.

Prakash et al. [14] used multiple linear regression, support vector regression, and recurrent neural network methods to estimate soil moisture for 1,2 and 7 days in three different datasets (with 569, 4749 and 92 samples) collected from different online pools. As a result of the evaluation, it was seen that the multiple linear regression is superior with the coefficient of determination (R^2) value of 0.975 for 1 day before, 0.939 for 2 days before and 0.786 for 7 days before.

Goap et al. [15] developed a Raspberry Pi and Arduino based system in the rose garden to collect data. In the experimental study, soil moisture sensor (VH-400), temperature and humidity sensor (DHT22) and Ultraviolet

(UV) sensor were used. Regression analysis was performed in R programming. Multiple linear regression (MLR), ridge regression, weighted linear regression and support vector regression (SVR) were created for the data obtained from the sensor. SVR found the best result with 0.9383 R value.

Sing et al. [16] developed an application to optimize the use of water by estimating the soil moisture value. Air temperature, air humidity, soil moisture, soil temperature, radiation sensors were installed to collect data from the IoT-based smart irrigation system and also weather forecast data from the Internet was used to estimate the soil moisture value. Machine learning techniques such as Gradient Boosted Regression Tree (GBRT), Random Forest (RF), MLR and Elastic Net Regression (ENR) were used for prediction. The results show that GBRT outperforms the other methods with R^2 of 0.94.

When it examined the studies given above MLR, SVR, neural networks are the commonly used methods in soil moisture estimation. In the addition to the mostly evaluated methods in the literature, in this paper the performance of polynomial regression (PR), decision tree regression (DT) and RF algorithm are also evaluated.

The rest of the paper is organized as follows. In section 2, system design architecture and machine learning algorithms are given. In section 3, the results are shown. Finally, discussion and conclusions are given in section 4.

2. Material and Method

Machine Learning is the automatic detection of meaningful patterns in data [17]. There are various machine learning algorithms such as regression, classification, clustering, neural networks regarding the properties of the problems. In this section, the regression methods used in the study are briefly introduced.

2.1. Machine Learning Regression Algorithms

Multiple Linear Regression; MLR models the linear relationship between the dependent (target) variable with a numerical value and a set of independent variables. Thanks to this defined relationship, all information about independent variables can be found and this information can be used to produce more effective and accurate estimates. Its mathematical representation is given in Eq.1.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_N X_N + \varepsilon \tag{1}$$

Polynomial Regression; PR is a method that tries to explain such a relationship when there is a nonlinear relationship between data. Its mathematical representation is given in Eq.2.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2 + \beta_3 X_1^2 \ldots\ldots + \beta_N X_N^N + \varepsilon \tag{2}$$

N is the degree of polynomial here. Nonlinear connections are expressed between the independent variables used in the model.

Support Vector Regression**;** Before being used as a regression method, the support vector machine [18] used for classification aims to cover the maximum data point by keeping the margin range wide. This method, which has different kernel functions, allows us to draw curves suitable for our data set. Mathematical representation of linear SVR is given in Eq.3 and nonlinear SVR is given in Eq.4.

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*). < x_i, x > + b \tag{3}$$

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*). < Q(x_i), Q(x) > + b \tag{4}$$

Decision Tree Regression; It is the regression method that makes predictions using the tree structure [19]. In the structure created according to the data points, the decision variables are found in the nodes, and the values found by calculating the region averages as a result of the forecast are found in the leaves. The number of leaf nodes gives the dependent variables produced by estimation. The prediction process starts from the root node, tests the decision variables and continues until the leaf reaches the nodes. The last leaf node reached by the method is the result of the node estimation.

Random Forest Regression; Random forest regression generates multiple decision trees on a randomly selected dataset. The arguments that are input for the prediction process move through each tree structure created, reach the leaf node and generate the value. It produces predictive value by taking the average of the leaf node values found by each tree [20].

2.1. System Design Architecture

The block diagram of the system used to collect data from the IoT based irrigation system is given in Figure 1. The data received from the sensor node in the system is read by Arduino and saved to the cloud with the Esp8266 Wifi module.

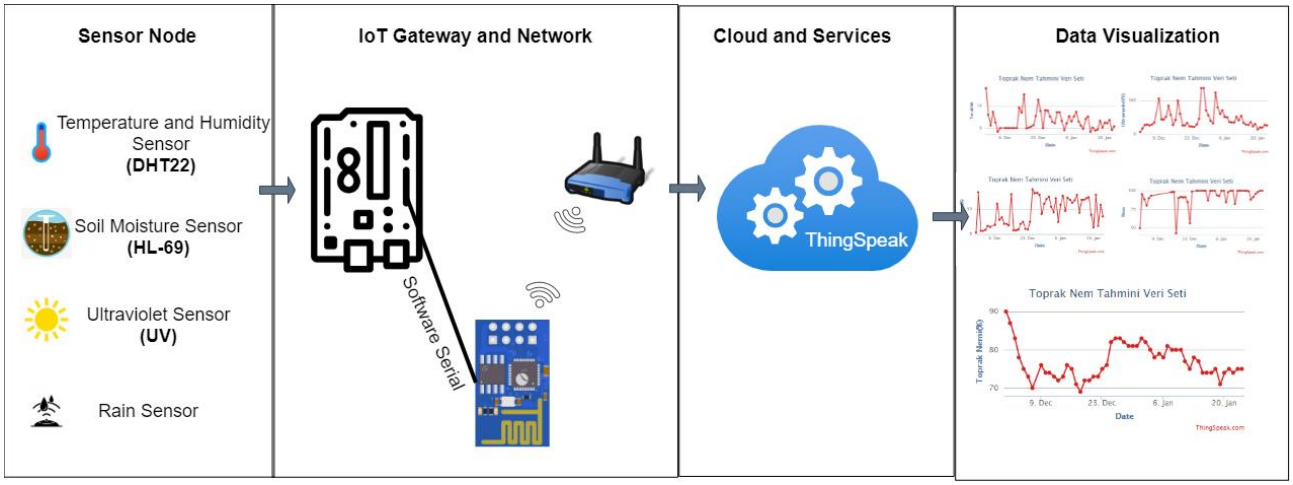

Figure 1. Block Diagram of the System

The environmental data taken from the sensors are exported to the internet and stored in the cloud system. ThingSpeak API (Application Programming Interface) is used to save data in the cloud system. ThingSpeak API is an open source interface that listens to incoming data, includes time information and outputs for humans through visual graphics and for machines through easily parsable code [21]. Esp8266 Wifi module sends data received from sensors to ThingSpeak. The ThingSpeak channel created to record the data received from the prototype system is given in Figure 2. In the soil moisture estimation data set channel, 5 areas are created to record the data from the system.
The key of the channel is used to write the data to the correct channel. In addition to storing data, ThingSpeak offers visualization and remote monitoring of these data over the internet.

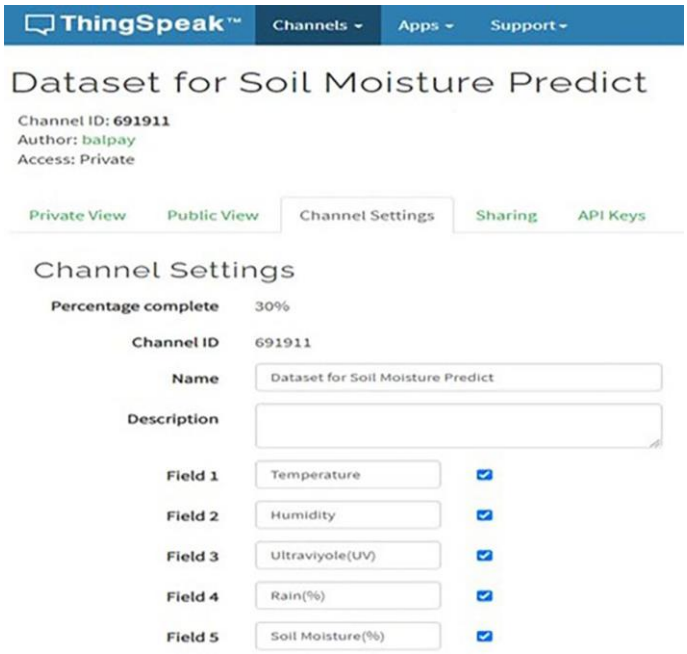

Figure 2. ThingSpeak Channel Display

An HTTP request is sent by the system to the ThingSpeak channel with the Esp8266 Wifi Module. Within this request, there is a channel-specific write key and data of 5 fields read from sensors via Arduino. During 55 days, data has been taken from the system every half hour and recorded in ThingSpeak. The data recorded in the channel is taken in .csv format and used for training and testing the models. In the data set containing 2395 samples in total, the data received as "NAN" when the sensor could not read is completed by using the average strategy in the preprocessing stage.

As the soil moisture parameter is affected by the momentary environmental conditions, the soil moisture information of the previous day is also important. For this reason, the average soil moisture information of the previous day is included in the data set.

2.2. Data preprocessing

Since the statistical properties of the attributes in the data set are different, some attributes may override others while the model is being formed. In this case, there are 2 methods in data science as normalization and standardization to reduce the attributes to the same plane. In normalization, the smallest value is set to 0 and the largest value is set to 1. In standardization, the mean value (μ) is set to zero. In this study, the min-max normalization method calculated by Eq.5, in which the values are spread between 0 and 1, has been used as a pre-process. The data set has been randomly divided into 75% training and 25% testing for use in experimental studies. The calculations are performed in the Python programming language compiler Spyder. Machine learning library scikit-learn is used for normalization, training and testing data creation, and running regression algorithms.

$$X' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{5}$$

3. Results

In this study, machine learning algorithms MLR, PR, DT, SVR and RF regression methods are used. The models of these methods are run using the scikit-learn library. While training the regression models in the experiments, temperature, humidity, UV, rain and the previous day's average soil moisture parameter data collected from the field are used. The success of the models trained on the basis of R^2 and MSE values are evaluated. These two criteria basically indicate how close the estimates produced by the models are to the real values. The proximity of the value to 1 in the R^2 method and the value to 0 in the MSE method show the success of the regression models. Predictions are made by giving test data as input to the regression models created as a result of the training. The graph showing the estimated and actual soil moisture values of the MLR model is given in Figure 3.

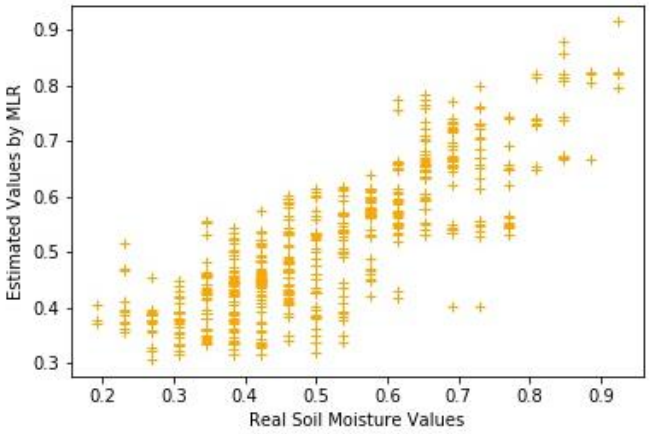

Figure 3. Multiple linear regression values

Figure 4 shows the graph showing the predicted values of the PR model. While creating the polynomial regression model, the polynomial degree is calculated as 2.

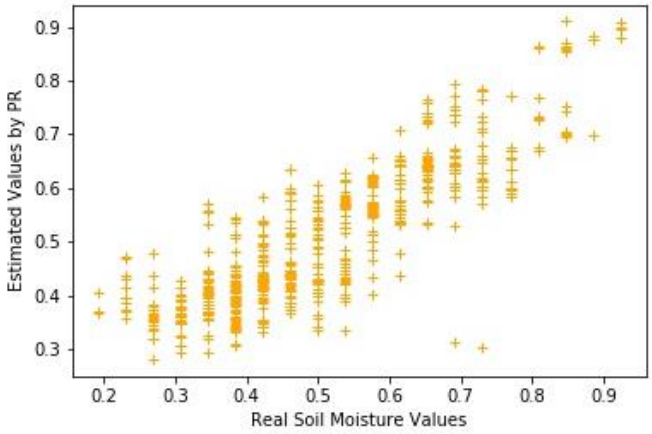

Figure 4. Polynomial regression values

The graph of the SVR model using 'rbf' as the kernel function is given in Figure 5.

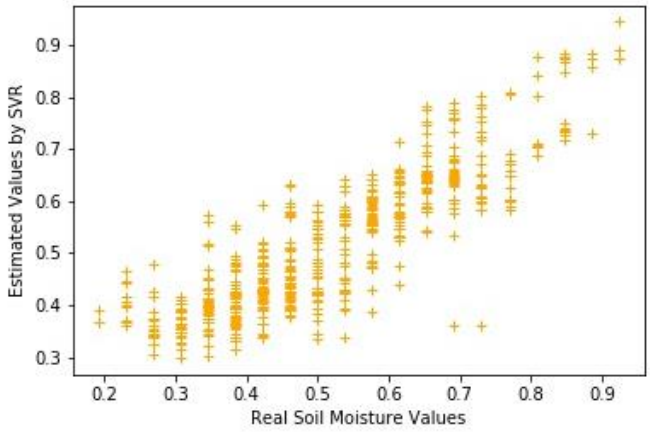

Figure 5. Support vector regression values

Figure 6 shows the graph showing the predicted values of the DT model. In the DT regression model, the maximum depth parameter of the tree is taken as 5.

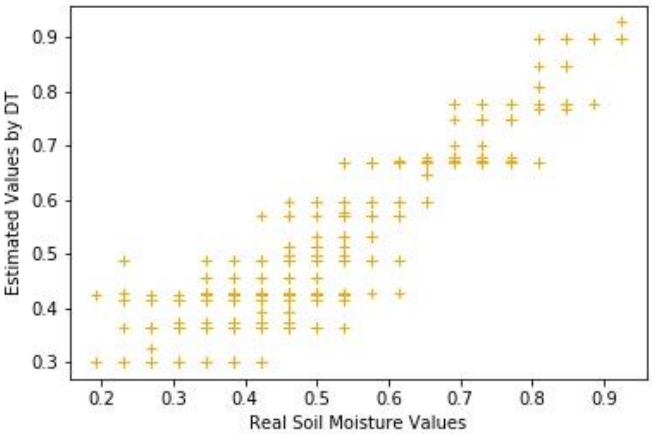

Figure 6. Decision tree regression values

Figure 7 shows the graph showing the predicted values of the RF model.

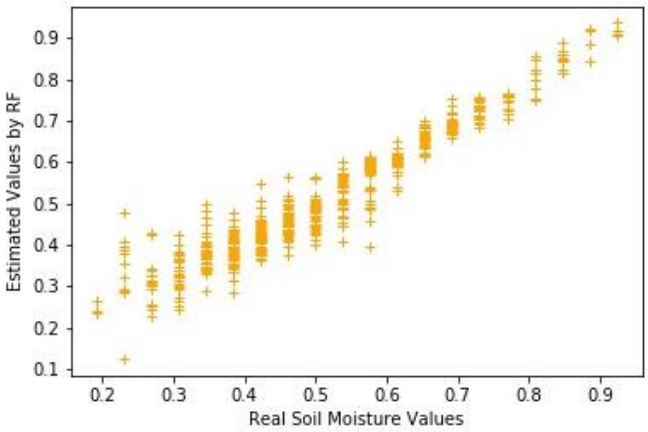

Figure 7. Random forest regression values

Figure 3-7 shows the values predicted by each regression algorithm against the real values in the test data. Since the real values do not match the predicted values in Figure 3, which belongs to the MLR, the spread of the graph over a wide area is seen. This makes it difficult to establish a trendline for the model, and the margin for error is high. While creating the PR model, the polynomial degree was taken as 2 and a graph with a narrower spread was obtained compared to the MLR. SVR model shown in Figure 5, on the other hand, obtained better predictions than MLR and PR, since the outputs of DT model in Figure 6 are the values calculated in leaf node, the model gives certain discrete outputs in the predictions. A graded graphic has been observed and the performance of the model is better than MLR, PR and SVR. In the RF model shown in Figure 7, it has been observed that the actual and predicted values coincide. In addition, when Figure 3-7 is examined independently from R^2 and MSE values, it is seen that the actual and predicted values in Figure 7 have a trend line and produce better results than other algorithms. R^2 and MSE values calculated according to the real and predicted values on the test set of MLR, PR, DT, SVR and RF models that estimate the soil moisture value are given in Table 1.

Table 1. R^2 and MSE values of the algorithms

Regression Algorithms	R^2	MSE
Multiple Linear Regression	0.681	0.0077
Polynomial Regression	0.730	0.0065
Support Vector Regression	0.751	0.0060
Decision Tree Regression	0.836	0.0039
Random Forest Regression	0.926	0.0018

4. Discussion and Conclusion

Soil moisture is a critical parameter used to develop smart irrigation systems, measure soil fertility, monitor droughts, increase crop yields, improve weather forecasts, predict floods, and has a complex relationship with environmental factors. In this study, models that predict the soil moisture value, which has an important place in agricultural practices, have been presented. The success of these models has been compared according to R^2 and MSE criteria. During the training and testing of the models, the data obtained from the system developed for 55 days are used. MLR, PR, SVR, DT and RF regression algorithms have been applied for estimation. In the RF regression model, R^2 and MSE values have been calculated as 0.926 and 0.0018, respectively, and this model has been found to be more successful than other models. Here, the failure of the multiple linear regression method shows that there is no linear relationship between the attributes in the estimation of the soil moisture value. The fact that tree-based methods are more successful is that they can discover complex relationships between attributes, and that the variable value range is narrower because the data set used in the study is collected in a short period.

As a future study, the effect of enriching the dataset by using more sensor nodes and using different data preprocessing steps on the performance of the methods will be analyzed. Also, with the advancement in technology, weather forecast accuracy has improved significantly. Weather forecast data can be used to predict changes in soil moisture.

References

[1] Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., Sheth, A. P. 2018. Machine learning for internet of things data analysis: a survey. Digital Communications and Networks, 4(3), 161–175.

[2] Hassija, V., Chamola, V., Saxena, V., Jain, D., Goyal, P., Sikdar, B. 2019. A Survey on IoT Security: Application Areas, Security Threats, and Solution Architectures. IEEE Access, 7, 82721–82743.

[3] Pernapati, K. 2018. IoT Based Low Cost Smart Irrigation System. Proceedings of the International Conference on Inventive Communication and Computational Technologies, April, India, 1312–1315.

[4] Thakare, S., Bhagat, P. H. 2019. Arduino-Based Smart Irrigation Using Sensors and ESP8266 WiFi Module. Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, June, India, 1085–1089.

[5] Balamurugan, C., Satheesh, R. 2017. Development of Raspberry pi and IoT Based Monitoring and Controlling Devices for Agriculture. Journal of Social, Technological and Environmental Science, 6, 207–215.

[6] Capello, F., Toja, M., Trapani, N. 2016. A real-Time monitoring service based on industrial internet of things to manage agrifood logistics. ILS 2016 - 6th International Conference on Information Systems, Logistics and Supply Chain, 1–8 June, France, 1-8.

[7] Chen, K. T., Zhang, H. H., Wu, T. T., Hu, J., Zhai, C. Y., Wang, D. 2014. Design of monitoring system for multilayer soil temperature and moisture based on WSN. Proceedings - 2014 International Conference on Wireless Communication and Sensor Network, November, India, 425–430.

[8] Minbo, L., Zhu, Z., Guangyu, C. 2013. Information Service System of Agriculture IoT. Automatika, 54(4), 415–426.

[9] Payero, J. O., Mirzakhani-Nafchi, A., Khalilian, A., Qiao, X., Davis, R. 2017. Development of a Low-Cost Internet-of-Things (IoT) System for Monitoring Soil Water Potential Using Watermark 200SS Sensors. Advances in Internet of Things, 07(03), 71–86.

[10] Liu, J. 2016. Design and Implementation of an Intelligent Environmental-Control System: Perception, Network, and Application with Fused Data Collected from Multiple Sensors in a Greenhouse at Jiangsu, China. International Journal of Distributed Sensor Networks, 12(7), 1-10.

[11] Yang, J., Liu, M., Lu, J., Miao, Y., Hossain, M. A., Alhamid, M. F. 2018. Botanical Internet of Things: Toward Smart Indoor Farming by Connecting People, Plant, Data and Clouds. Mobile Networks and Applications, 23(2), 188–202.

[12] Shekhar, Y., Dagur, E., Mishra, S., Tom, R. J., Veeramanikandan, M., Sankaranarayanan, S. 2017. Intelligent IoT Based Automated Irrigation System. International Journal of Applied Engineering Research, 12(18), 7306–7320.

[13] Gorthi, S., Dou, H. 2011. Prediction Models for the Estimation of Soil Moisture Content. Proceedings of the ASME 2011 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, 1–9 August, US, 1-9.

[14] Prakash, S., Sharma, A., Sahu, S. S. 2018. Soil Moisture Prediction Using Machine Learning. Proceedings of the International Conference on Inventive Communication and Computational Technologies, 20-21 April, India, 1–6.

[15] Goap, A., Sharma, D., Shukla, A. K., Krishna, C. R. 2018. Comparative Study of Regression Models Towards Performance Estimation in Soil Moisture Prediction. International Conference on Advances in Computing and Data Sciences, April, India, 309-316.

[16] Singh, G., Sharma, D., Goap, A., Sehgal, S., Shukla, A. K., & Kumar, S. 2019. Machine Learning based soil moisture prediction for Internet of Things based Smart Irrigation System. Proceedings of IEEE International Conference on Signal Processing, Computing and Control, October, 175–180.

[17] Shalev-Shwartz, S., Ben-David, S. 2013. Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press.

[18] Vapnik, V., Lerner, A. 1963. Generalized Portrait Method for Pattern Recognition. Automation and Remote Control, 24(6), 774–780.

[19] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, C. J. S. 1984. Classification and Regression Trees. In Chapman and Hall/CRC.

[20] Breiman, L. 1996. Bagging Predictors. Machine Learning, 24(0), 123–140.

[21] Gómez Maureira, M. A., Oldenhof, D., & Teernstra, L. 2014. ThingSpeak – an API and Web Service for the Internet of Things. https://staas.home.xs4all.nl/t/swtr/documents/wt2014_thingspeak.pdf, Accessed 20 April 2021.