DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

# DEVELOPMENT OF INTERNET OF THINGS (IoT) BASED SYSTEM FOR AGRICULTURE USING MACHINE LEARNING METHODS

by

Cansel KÜÇÜK

June, 2021

İZMİR

# DEVELOPMENT OF INTERNET OF THINGS (IoT) BASED SYSTEM FOR AGRICULTURE USING MACHINE LEARNING METHODS

**A Thesis Submitted to the**
**Graduate School of Natural and Applied Sciences of Dokuz Eylül University**
**In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering, Computer Engineering Program**

**by**

**Cansel KÜÇÜK**

**June, 2021**

**İZMİR**

## M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DEVELOPMENT OF INTERNET OF THINGS (IoT) BASED SYSTEM FOR AGRICULTURE USING MACHINE LEARNING METHODS"** completed by **CANSEL KÜÇÜK** under supervision of **ASSOC. PROF. DR. DERYA BİRANT** and **ASSIST. PROF. DR. PELİN YILDIRIM TAŞER** and we certify that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

| ……………………………………….. | ……...…………………………………… |
|---|---|
| Assoc. Prof. Dr. Derya BİRANT | Assist. Prof. Dr. Pelin YILDIRIM TAŞER |
| Supervisor | Supervisor |

| ……………………………………….. | ….………………………………… |
|---|---|
| Assist. Prof. Dr. Özlem AKTAŞ | Assist. Prof. Dr. Fatma BOZYİĞİT |
| (Jury Member) | (Jury Member) |

Prof. Dr. Özgür ÖZÇELİK

Director

Graduate School of Natural and Applied Sciences

## ACKNOWLEDGMENTS

First of all, I would like to thank my advisors, Assoc. Prof. Dr. Derya BİRANT and Assist. Prof. Dr. Pelin YILDIRIM TAŞER, for their supervision and invaluable suggestions through this study. Their leadership helped me during the research and the writing of this thesis.

I thank Volkan KIRANOĞLU and Göksu TÜYSÜZOĞLU for supporting the preparation of the dataset used in this study. I would also like to extend my special thanks to Izmir Metropolitan Municipality Agricultural Engineer Ali Selim ALPASLAN for enabling us to use Izmir Municipality's data.

I would also like to express my gratitude to my parents and my twin sister who have always believed in me throughout my life, and who have always supported me spiritually while preparing this thesis. Also, I would like to thank my business partner and best friends.

Cansel KÜÇÜK

# DEVELOPMENT OF INTERNET OF THINGS (IoT) BASED SYSTEM FOR AGRICULTURE USING MACHINE LEARNING METHODS

## ABSTRACT

Internet of Things (IoT) has become one of the common technologies used in the agriculture field. The prediction of soil temperature and soil moisture in agricultural field has a significant role in the growth and development of plants. Considering this motivation, in this thesis, two case studies were performed for predicting soil temperature and soil moisture on real-world datasets collected by IoT sensors.

In the first study, a novel method, named *Soil Temperature Ordinal Classification* (STOC), which considers the relationships between the class labels (i.e., low, medium, high) during soil temperature prediction was proposed. To prove the effectiveness of the proposed approach, we applied the STOC method using five different machine learning algorithms (decision tree, Naive Bayes, k-nearest neighbors, support vector machines, and random forest) to a real-world data obtained by IoT sensors from 16 stations in three states (Utah, Alabama, and New Mexico) of United States at five soil depths (2, 4, 8, 20, and 40 inches) between the years of 2011 and 2020.

In the second study, an intelligent *Multi-Output Regression for Soil Moisture Prediction* (MOR-SMP) method was implemented for estimating soil moisture at three soil depths (15, 30, and 45 cm). This approach was tested by applying nine different machine learning algorithms on daily values of meteorological and soil data obtained by IoT sensors from Kemalpaşa-Örnekköy station in Izmir, Turkey.

The experimental results showed that the proposed STOC and MOR-SMP methods achieved good performance in predicting soil temperature and soil moisture, respectively. Hence, they can be effectively used in the agriculture sector.

**Keywords:** Machine learning, classification and regression, agriculture, soil temperature prediction, soil moisture prediction, internet of things

**MAKİNE ÖĞRENMESİ YÖNTEMLERİNİ KULLANARAK TARIM İÇİN NESNELERİN İNTERNETİ (IoT) TABANLI SİSTEM GELİŞTİRİLMESİ**

## ÖZ


Nesnelerin İnterneti (IoT), tarım alanında kullanılan yaygın teknolojilerden biri haline gelmiştir. Tarım arazilerindeki toprak sıcaklığı ve toprak neminin tahmini, bitkilerin büyümesi ve gelişmesinde önemli bir role sahiptir. Bu motivasyon göz önünde bulundurularak, bu tezde, IoT sensörleri tarafından toplanan gerçek veri setlerinden toprak sıcaklığını ve toprak nemini tahmin etmek için iki çalışma yapılmıştır.

İlk çalışmada, toprak sıcaklığı tahmini sırasında sınıf etiketleri arasındaki ilişkileri (örneğin, düşük, orta, yüksek) dikkate alan, *Toprak Sıcaklığı Sıralı Sınıflandırması* (STOC) adlı yeni bir yöntem önerilmiştir. Önerilen yaklaşımın etkinliğini göstermek için, beş farklı makine öğrenmesi algoritması (karar ağaçları, Naive Bayes, k-en yakın komşu, destek vektör makineleri ve rastgele orman) kullanarak STOC yöntemi; Amerika Birleşik Devletleri'nin üç eyaletindeki (Utah, Alabama, ve New Mexico) 16 istasyondan IoT sensörleri tarafından 2011 ve 2020 yılları arasında beş toprak derinliğinden (2, 4, 8, 20, ve 40 inç) elde edilen gerçek verilere uygulanmıştır.

İkinci çalışmada, üç toprak derinliğinde (15, 30, ve 45 cm) toprak nemini tahmin etmek için akıllı *Toprak Nemi Tahmini için Çok-Çıkışlı Regresyon* (MOR-SMP) yöntemi uygulanmıştır. Bu yaklaşım, Türkiye İzmir Kemalpaşa-Örnekköy istasyonundan IoT sensörleri tarafından elde edilen toprak verileri ve günlük meteorolojik veriler üzerine dokuz farklı makine öğrenmesi algoritması kullanılarak test edilmiştir.

Deneysel sonuçlar, önerilen STOC ve MOR-SMP yöntemlerinin sırasıyla toprak sıcaklığını ve toprak nemini tahmin etmede iyi performans verdiğini göstermiştir. Böylelikle, önerilen yöntemler tarım sektörü için etkin bir şekilde kullanılabilecektir.

**Anahtar Kelimeler:** Makine öğrenmesi, sınıflandırma ve regresyon, tarım, toprak sıcaklığı tahmini, toprak nemi tahmini, nesnelerin interneti

**CONTENTS**

**Page**

**LIST OF FIGURES**

**Page**

**LIST OF TABLES**

**Page**

# CHAPTER ONE
# INTRODUCTION

## 1.1 General

The Internet of Things (IoT) devices contain sensors and mini-computer processors that act on the sensors' data, and it advanced from machine-to-machine (M2M) correspondence, i.e., machines associating by means of an organization without human cooperation. Nowadays, IoT technology has been used in many areas, ranging from marketing to automation and banking to education. Many devices used in daily life contain IoT technology such as white goods, street lamps, mobile phones, car parking, coffee and tea maker, and in many items, we use daily. Hereafter, it is thought that many new products will include IoT technology.

Wearable devices consisting of specific software and sensors can analyze data collected from people and communicate with different devices. Especially in healthcare, IoT monitors patients' symptoms and gives warnings and provides the opportunity to follow any sudden reaction more closely by analyzing the collected data. In addition, smart buildings can reduce unnecessary expenses by measuring water and electricity consumption costs by taking into account the number of people in a house and giving warnings. In smart cities, energy consumption can be reduced with smart meters and smart street lights, and IoT sensors can help improve environmental concerns.

Farming and agriculture have seen many technological transformations in the last decades, becoming more industrialized and technology-driven. Various machines, animals and plants located at different points in the agricultural land with IoT devices interact with each other. IoT technology has started to be preferred in the agriculture field. For example, with the help of sensors used with IoT systems in agriculture, field light can be measured and soil temperature and humidity can be monitored easily.

According to Turkish Statistical Institute (TUİK) data, about 16% of the working population of our country is employed in agriculture (Karakaş, 2020). Turkey is a country that has quite convenient land, where the total agricultural area is 37753 thousand hectares (Tarım ve Orman Bakanlığı [TOB], 2020). Hence, it is important to make estimations with the data obtained from the soil in order to facilitate production in agriculture and obtain lower-cost agricultural products. Therefore, soil temperature and soil moisture, which are quite important parameters for agriculture efficiency, can be predicted using machine learning techniques.

### 1.2 Purpose

*Soil temperature* has a great influence on plant growth and development, soil water and salt transport, soil carbon balance, and microbial activity and chemical reactions inside the soil (Onwuka & Mang, 2018). Biochemical processes, such as germination of a seed, seedling emergence, uptake operation by roots, root growth, are realized under the suitable soil temperature. The high temperature in the soil can lead to a dramatic reduction of soil resistance to physical events such as erosion and subsidence. On the other hand, when the temperature drops dramatically and the soil freezes, some biochemical activities can stop. In addition, a significant decrease in temperature changes the rate of organic matter decomposition and mineralization inside the soil. If the soil temperature is suitable, chemical activities in the soil continue smoothly and regularly. Soil temperature estimation is of great importance for this purpose. If we predict further changes in soil temperature, we can develop new strategies to prevent undesired situations.

*Soil moisture* is also another important factor for plant diversity, crop yield, sustenance of any agricultural system, thermal properties of soils, and estimating flood, slope failure, and erosion (Sanuade et al., 2020). Monitoring and evaluating soil moisture can be costly because of the sensors and their regular maintenance. Because of this reason, soil moisture prediction in advance supports the decision-making process in the field of agriculture, such as the management of water reservoirs for better crop yield, cost-saving, and plant diversity. In most cases, the soil moisture has been estimated using traditional methods in a laboratory environment. However, these techniques can be time-consuming with huge datasets and also cannot discover hidden patterns in complex structural properties of meteorological and soil data (Kucuk et al., 2021).

Considering these motivations, in the last decade, *machine learning* techniques have been used in the agriculture field to make more accurate soil temperature and soil moisture predictions. In this thesis, two different machine learning studies were performed on real-world meteorological and soil data to obtain valuable knowledge and make the right decisions in the field of agriculture. The purposes of these studies are listed below:

- Predicting soil temperature of soil data at five soil depths (2, 4, 8, 20, and 40 inches) using an ordinal classification approach that considers the relationships between the class labels (i.e., low, medium, high).
- Predicting soil moisture at three soil depths (15, 30, and 45 cm) using an intelligent multi-output regression method.

### 1.3 Main Contributions of the Thesis

The main contributions of this thesis are on two folds.

First, we proposed a new ordinal classification approach: *Soil Temperature Ordinal Classification* (STOC) which considers the relationships between the class labels (i.e., low, medium, high) during soil temperature prediction. To demonstrate the effectiveness of the proposed approach, we applied the STOC method using five different traditional machine learning algorithms (decision tree, Naive Bayes, k-nearest neighbors, support vector machines, and random forest) on daily values of meteorological and soil data obtained from 16 stations in three states (Utah, Alabama, and New Mexico) of United States at five soil depths (2, 4, 8, 20, and 40 inches) between the years of 2011 and 2020. The experiments show that the proposed STOC approach is an efficient method for soil temperature prediction. All of the applied STOC models, except STOC.NB, reached over 75.5% soil temperature prediction performance. Also, it was observed from the experimental results that the STOC.DT achieved the best soil temperature estimation with an average accuracy rate of 90.95%.

Second, we implemented an intelligent *Multi-Output Regression for Soil Moisture Prediction* (MOR-SMP) method. We applied the method on daily values of meteorological and soil data obtained from Kemalpaşa-Örnekköy station in Izmir, Turkey at three soil depths (15, 30, and 45 cm) between the years of 2017 and 2019. In this study, nine different machine learning algorithms (Linear Regression (LR), Ridge Regression (RR), Least Absolute Shrinkage and Selection Operator (Lasso), Random Forest (RF), Extra Tree Regression (ETR), Adaptive Boosting (AdaBoost), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), and Histogram-Based Gradient Boosting (HGB)) were compared each other in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination ($R^2$) metrics. The experiments demonstrated that the multi-output regression model showed good soil moisture prediction performance. According to the results, the ETR algorithm achieved the best prediction performance with an 0.81 $R^2$ value among the other models.

### 1.4 Organization of the Thesis

The thesis consists of seven chapters. The remainder of this thesis is structured as follows:

In Chapter 2, previous studies about soil temperature prediction and soil moisture prediction are summarized. Besides, a comparison of previous studies is presented.

In Chapter 3, the ordinal classification and multi-output regression tasks are described in detail.

In Chapter 4, a novel approach, named STOC, is explained with its formal definition. After that, some information about the five different traditional machine learning methods (decision tree, Naive Bayes, k-nearest neighbors, support vector machines, and random forest) is given.

In Chapter 5, an intelligent method, called MOR-SPM, is presented with its formal definition. In addition, this chapter also gives brief information about the nine different machine learning algorithms (LR, RR, Lasso, RF, ETR, AdaBoost, GB, XGBoost, and HGB).

In Chapter 6, the experimental studies are explained with the dataset description and the obtained experimental results.

In the last chapter, some concluding remarks and future directions are presented.

# CHAPTER TWO
# LITERATURE REVIEW

## 2.1 Review of Previous Studies

Last years, there has been a worldwide rise in agricultural data analysis. Some studies have been focused on the prediction of soil temperature and soil moisture. The following is a summary of the relevant studies.

### *2.1.1 Review of Soil Temperature Prediction*

*Machine learning* has recently received much attention in the soil temperature prediction field (Sattari et al., 2020; Li et al., 2020; Abyaneh et al., 2016). It discovers significant underlying patterns in raw data by constructing a model without any human intervention. Classification is the most widely studied machine learning technique *Classification* is the task of categorizing an input sample data into one of the predefined classes. In the literature, there are several soil temperature prediction studies that apply different machine learning techniques, such as regression (Alizamir et al., 2020a; Alizamir et al., 2020b; Sattari et al., 2020; Abyaneh et al., 2016; Li et al., 2020) and time series (Zeynoddin et al., 2020; Mehdizadeh et al., 2020). Such as, Alizamir et al., (2020b) proposed a deep echo state network (Deep ESN) regression model for soil temperature forecasting at 10 and 20 cm depths. In another study, a new time series model, called fractionally autoregressive integrated moving average (FARIMA), was built for daily soil temperature prediction at four depths (5, 10, 50, and 100 cm) (Mehdizadeh et al., 2020).

Some soil temperature forecasting studies were performed at multiple-depths (Kisi & Tombul, 2015; Sanikhani et al., 2018; Shamshirband et al., 2020). For example, Sanikhani et al., (2018) applied extreme learning machine (ELM), artificial neural network (ANN), and M5 Model Tree (M5 Tree) models on the meteorological data obtained from two stations in Turkey for predicting soil temperatures at 5, 50, and 100 cm depths. Some studies investigated the features that affect the soil temperature prediction (Feng et al., 2019; Nanda et al., 2020). Nanda et al. (2020) discovered that rainfall data does not affect the prediction performance so much, while the soil moisture parameter significantly improves the accuracy of the prediction problem.

In the machine learning-based soil temperature prediction studies, it is assumed that there is no order between the class labels of the dataset which will be predicted. However, the class attributes of some soil data have an inherent order. To overcome this limitation, in this study, the ordinal classification technique which considers natural order between class labels was used for solving the soil temperature prediction problem.

### *2.1.2 Review of Soil Moisture Prediction*

In the literature, there are several studies (Matei et al., 2017; Singh et al., 2019) that used several machine learning algorithms for predicting soil moisture such as support vector machine (SVM), decision tree (DT), and logistic regression (LoR).

Nowadays, deep learning (DL) has been an active paradigm in soil moisture prediction studies to handle a large amount of meteorological and soil data (Yamaç et al., 2020; Cai et al., 2019). Yamaç, Şeker & Negiş (2020) implemented deep learning (DL), neural network (NN), and k-nearest neighbors (KNN) models for estimating moisture of calcareous soil data collected from Konya-Çumra plain, Turkey. Deep learning has the ability to solve high degree of complex problems rapidly and accurately. For this reason, it has been preferred to increase classification accuracy and reduce the margin of error in regression problems.

Researchers in another study (Cai et al., 2019) proposed a deep learning regression network (DNNR) model for predicting soil moisture in the test area located in Beijing, China. They implemented the DNNR model because of its advantages such that it can correlate or discover feature combinations, and is good at fusing hidden feature attributes, reducing the complexity of feature engineering, and improving the generalization capability of the model (Cai et al., 2019).

The k-nearest neighbors (KNN) algorithm has also been widely used to analyze agricultural data since it is easy to implement and at the same time simple, efficient, intuitive, and competitive (Yamaç et al., 2020).

Hong, Kalbarczyk & Iyer (2016) used the relevance vector machine (RVM) and the support vector machine (SVM) techniques for analyzing the agriculture data that was collected by sensors. They made predictions over the soil and climatic conditions.

As a result of their observations in their studies, Sanuade and his colleagues concluded that more effort and time should be spent in determining the content of soil moisture. They envisioned offering an alternative method to estimate soil moisture. Using three machine learning algorithms (Support Vector Machine, Fuzzy Logic and Artificial Neural Network), they reached an acceptable error rate of 5.65% on average. Based on this result, they concluded that using machine learning methods in the field of agriculture would yield positive results such as reducing costs and spending less time and effort (Sanuade et al., 2020).

Prasad, Deo & Maraseni (2018) built on the hybrid model in their work. They tested full ensemble empirical mode decomposition (CEEMDAN) and community empirical mode decomposition (EEMD) techniques, which are computationally efficient and self-adaptive multiresolution algorithms. Unlike other studies, it was integrated with multi-resolution vehicles (ELM) model and transformed into hybrid CEEMDAN-ELM and EEMD-ELM models and compared with random forest (RM) equivalent models.

Regression and deep learning methods in machine learning have been mostly used for soil moisture estimation. However, in the previous studies, these methods have been utilized to estimate a single target in the foreground based on the given input feature set. In our study, the Multi-Output Regression (MOR) method was preferred to estimate multiple target attribute values simultaneously.

## 2.2 Comparison of Previous Studies

The comparison of our study with previous studies are shown in Table 2.1 and 2.2 for soil temperature and soil moisture prediction, respectively.

Table 2.1 shows the comparison of our STOC method with the existing studies. Our method differs from the existing methods in three respects. First, they performed the prediction of numeric target values, while we focus on the classification of ordinal and categorical target values. Second, we used different methods such as KNN and NB. Third, since their target values are numeric they used different evaluation metrics such as root mean squared error (RMSE), relative root mean squared error (RRMSE), etc., whereas we tested the performance with accuracy and F-score metrics.

Table 2.1 Comparison of our study with the existing studies (1=DT, 2=SVM, 3=NB, 4=KNN, 5=RF, 6=NN, 7=ELM, 8=Regression, T=Time Series, O=Ordinal Classification, MD=Multiple-Depth)

| Reference | Year | Algorithm | | | | | | | | Task | | MD | Country |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | T | O | | |
| Alizamir et al 2020a | 2020 | √ | | | | | √ | √ | √ | | | √ | Turkey |
| Alizamir et al 2020b | 2020 | √ | | | | √ | √ | | √ | | | √ | USA |
| Sattari et al 2020 | 2020 | √ | | | | | | | √ | | | √ | Turkey |
| Li et al 2020 | 2020 | | √ | | | √ | √ | | √ | | | | USA |
| Zeynoddin et al 2020 | 2020 | | | | | | | √ | | √ | | √ | Iran |
| Mehdizadeh et al 2020 | 2020 | | | | | | √ | | | √ | | √ | Iran |
| Nanda et al 2020 | 2020 | | √ | | | √ | √ | | | √ | | | India |
| Shamshirband et al 2020 | 2020 | | √ | | | | √ | | √ | | | √ | Iran |
| Feng et al 2019 | 2019 | | | | | √ | √ | √ | √ | | | | China |
| Sanikhani et al 2018; | 2018 | √ | | | | | √ | √ | √ | | | √ | Turkey |

Table 2.1 continues

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Abyaneh et al 2016 | 2016 | | | | | | √ | | √ | | | √ | Iran |
| Kisi & Tombul 2015 | 2015 | | | | | | √ | | √ | | | √ | Turkey |
| Proposed approach | | √ | √ | √ | √ | √ | √ | | | | √ | √ | USA |

Table 2.2 presents the comparison of the proposed MOR-SMP with the previous studies. Our study differs from the existing methods in two respects. First, they performed the prediction of only single output, while we focus on the prediction of multi-output values. Second, different models from the previous studies (LR, RR, Lasso, RF, ETR, AdaBoost, GB, XGBoost, and HGB) were implemented in this study.

Table 2.2 Comparison of our study with the previous studies (R=Reference, SO=Single-output, MO=Multi-Output)

| R | Year | Model | Prediction | | Country |
|---|---|---|---|---|---|
| | | | SO | MO | |
| Sanuade et al 2020 | 2020 | NN, Fuzzy Logic, SVM | √ | | Nigeria |
| Matei et al 2017 | 2017 | KNN, SVM, NN, LoR, DT, RF, LR | √ | | Romania |
| Prakash et al 2018 | 2018 | LR, SVM, NN | √ | | United States |
| Singh et al 2019 | 2019 | GB, Elastic Net, LR, RF | √ | | Not stated |
| Yamaç et al 2020 | 2020 | DL, NN, kNN | √ | | Turkey |
| Cai et al 2019 | 2019 | DL | √ | | China |

Table 2.2 continues

| | | | |
|---|---|---|---|
| Our Study | LR, RR, Lasso, RF, ETR, AdaBoost, GB, XGBoost, and HGB | √ | Turkey |

# CHAPTER THREE
# BACKGROUND INFORMATION

## 3.1 Ordinal Classification

*Ordinal classification* is a special type of classification which assumes the class values of the attributes in the dataset are ordered and finite (Frank & Hall, 2001).

While classifying, a categorical target is defined with the attribute value specific to each class. These are called classification algorithms. The categories representing each class have an order among each other. In addition, standard classification algorithms do not use sequence information and are kept as unordered values (Frank & Hall, 2001).

Measurement types are divided into four. These; interval, ratio, ordinal and nominal quantities. While ordinal and nominal differ from each other, interval and ordinal quantities are similar in terms of order. Their difference is that their values are measured in equal units and are constant. Temperature measurement can be given as an example of ratio quantities (Frank & Hall, 2001).

In Figure 3.1, Frank & Hall (2001) schematically showed how a standard classification is adapted to an ordinal estimation. They took the data fictionally from the temperature estimation problem. Also divided them into three classes as Cold, Mild, and Hot. While explaining the training process in the upper part, they explained the test process in the lower part.

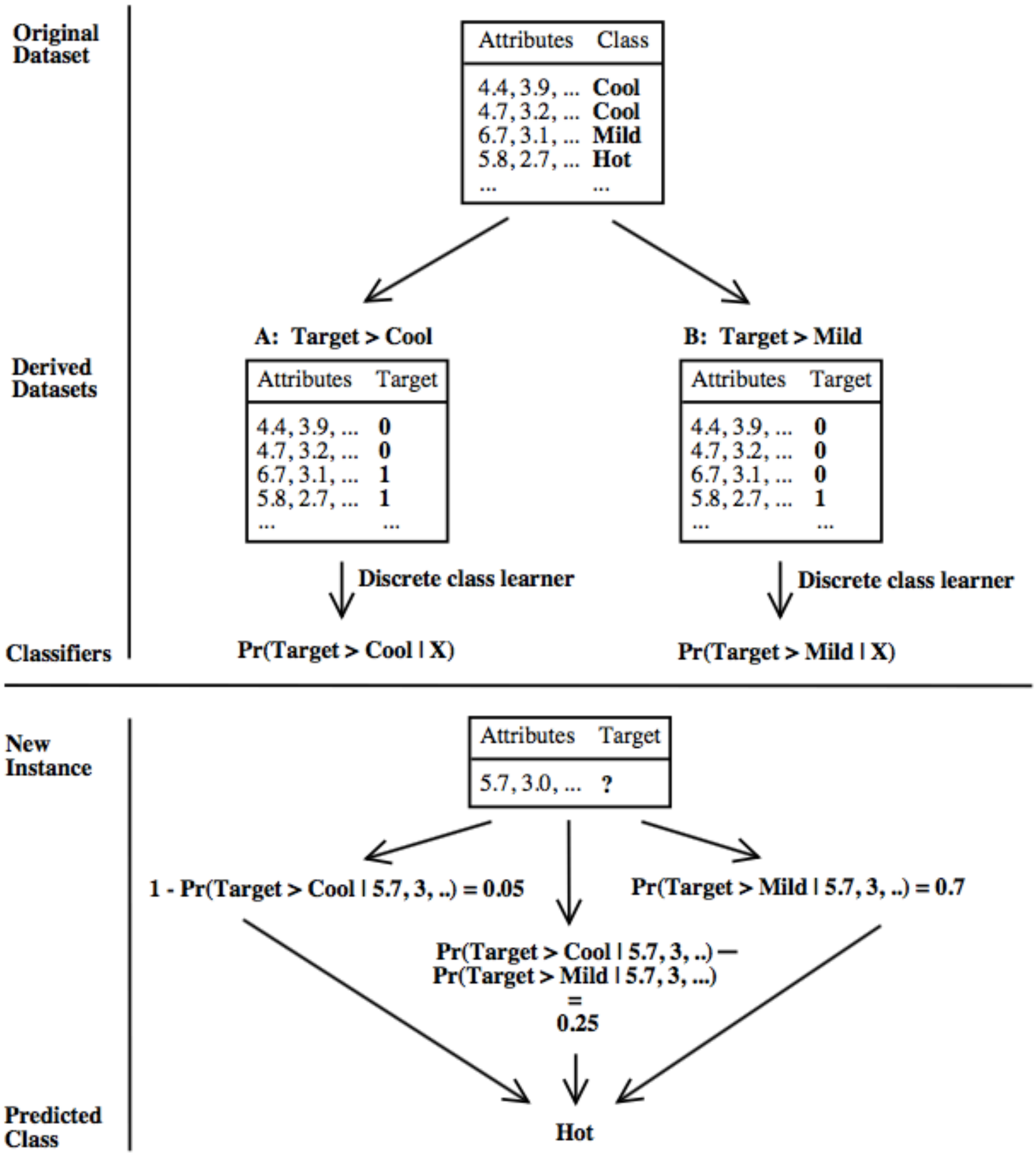


Figure 3.1 How standard classification algorithms are applied to ordinal prediction (Frank & Hall, 2001)

The aim of the ordinal classification is to predict the target attribute of a given new sample by considering ranking relationships between the classes. For example, the soil temperature values can be categorized as “very low”, “low”, “medium”, “high”, and “very high”. The order of these values is stated as “very low” < “low” < “medium” < “high” < “very high”. The prediction performance of the machine learning models can be improved regarding the ranking relationship between the class labels of the used experimental dataset. Considering this motivation, a novel ordinal classification method, named *Soil Temperature Ordinal Classification* (STOC), was applied to a real-world soil temperature data at five different soil depths (2, 4, 8, 20, and 40 inches) in this study.

### 3.2 Multi-Output Regression

*Multi-Output Regression* is a machine learning technique that predicts multiple target attribute values concurrently based on a given input attribute set.

With the single input feature given in multi-output regression, an output match with two or more variables is tried to be obtained. Previous algorithms defined by missing functions are similar to assuming there is Euclidean space. This is why the underlying structure of the output space is disregarded (Liu et al., 2009).

The system that expresses the situations in which a specified sample can be assigned to two or more classes is called multi-label learning. In the opinion of Liu et al., this is actually a situation that is observed very often in practice. It is especially common in music and text categories. For example; documents in the music category may belong to more than two genres, such as rock and blues, while in the text category documents may belong to genres such as government and medical. Again, especially in the field of medical, a disease can be in more than two categories. Multi-label and multi-class learning differ from each other in the aspect of assigning one class or more than one class for a given example (Liu et al., 2009).

Multi-output regression is one of the methods that has appeared recently. It has been observed that it has been used frequently in the field of computer vision and plays an essential role in the predictions made in this field. Zhen and his colleagues demonstrated the operativeness of the recommended Supervised Descriptor Learning (SDL) algorithm on the typical *Multi-Output Regression* task in their study with the data set they obtained from the visuals. They have carried out their work by setting a multivariate target for head pose estimation, which has been frequently studied recently. The SDL algorithm can obtain a compact and highly selective feature representation, which enables more precise and efficient multivariate prediction. The SDL algorithm is structured as a generalized low-rank approximation of matrices with a supervised manifold formalization built on a matrix. This also, obtains a compact while efficient dimensionality reduction algorithm (Zhen et al., 2015).

# CHAPTER FOUR
# PROPOSED APPROACH: SOIL TEMPERATURE ORDINAL CLASSIFICATION (STOC)

## 4.1 Ordinal Classification for Soil Temperature Prediction

This thesis proposes a new method, called *Soil Temperature Ordinal Classification* (STOC), which takes into account the relationships between the class labels (i.e., low, medium, high) during soil temperature prediction. This study is also original in that it compares alternative base learners in conjunction with the proposed method, including decision tree (DT), Naive Bayes (NB), k-nearest neighbors (KNN), support vector machines (SVM), and random forest (RF). This is the first study that estimates the soil temperatures at five different soil depths (2, 4, 8, 20, and 40 inches) by using ordinal classification.

Figure 4.1 shows a general overview of the proposed approach. First, in the data acquisition step, meteorological (air temperature, air pressure, precipitation, relative humidity, and solar radiation) and soil (soil temperature and soil moisture) data is obtained from the sensors (i.e. solar radiation sensor, wind speed, and direction sensor, etc.) of the Soil Climate Analysis Network (SCAN) in 16 stations of United States. These data are stored in a cloud platform. In the next step, first, the datasets are passed through a data preprocessing step (discretization and missing data elimination) and then they are converted to binary datasets to make them ready for the implementation of the STOC method. In the training phase, the STOC approach is applied to the meteorological and soil data using DT, NB, KNN, SVM, and RF as base learners. After that, the soil temperature estimation performances of these models are evaluated using the n-fold cross-validation technique selecting $n$ as 10. Then, in the prediction phase, a new sample is predicted using the STOC method. Finally, the obtained accuracy rate and F-measure results from these models are represented in the presentation step.

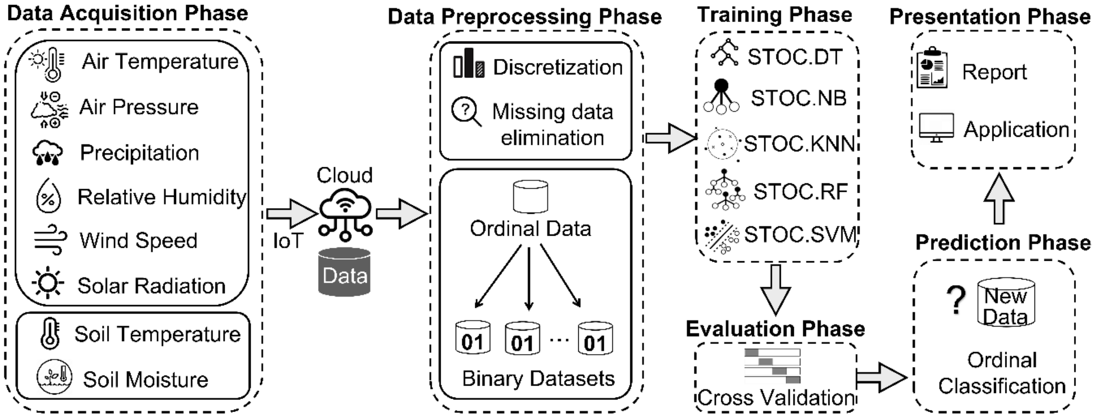


Figure 4.1 The general overview of the STOC approach

The main benefits of the proposed STOC approach can be summarized as follows:

• Using the meteorological data, future soil temperature values for a region can be forecasted.

• An intelligent soil management system can be designed using the predicted values.

• The crop yield in agriculture can be increased by using the predicted soil temperature values.

• The proposed approach can play a role in reducing unnecessary resource consumption in agriculture, such as water, sensor, and pesticide.

## 4.2 Formal Definition

**Definition 1**. *Soil temperature ordinal classification* refers to the problem that uses ordinal data obtained from sensors and aims to construct a prediction model by considering the orders among class labels to correctly estimate the level of the soil temperature according to the methodological data.

Let $D$ be an ordinal meteorological and soil data $D=\{(x_i, y_i)_{i=1}^{N} \epsilon\ X\ x\ Y\}$, where $X \epsilon R^d$ and its related class label $y_i$ belongs to the output space $Y=\{C_1, C_2, C_3, ...., C_k\}$ having $k$ classes with a data ranking structure $C_1 < C_2 < C_3 < .... < C_k$, where $<$ denotes the linear order relation. In this method, five different traditional classification algorithms (DT, NB, KNN, SVM, and RF) were implemented as base learners. The aim of classification is to categorize an input vector *x* with one of the *k* class labels by selecting the maximum probability value among the evaluated probabilities for each label using a binary classification method. In this approach, the probabilities for each *k* classes are computed as follows:

$$\begin{gathered} P(C_1 | x) = 1 - P(\text{Class} > C_1 | x) \\ P(C_j | x) = P(\text{Class} > C_{j-1} | x) - P(\text{Class} > C_j | x) \\ P(C_k | x) = P(\text{Class} > C_{k-1} | x) \end{gathered} \tag{4.1}$$

where $j=\{2,3,\ldots,k\}$.

### 4.3 Classification Methods

In this study, five different traditional classification algorithms (DT, NB, KNN, SVM, and RF) were chosen as base learners for the STOC method.

**Decision Tree (C4.5 Algorithm):** Decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles and the leaf nodes are denoted by ovals. C4.5 handles both categorical and continuous attributes to build a decision tree. It also handles missing attribute values (Yadav & Pal, 2012).

**Naive Bayes (NB):** NB is a simple learning algorithm. In this learning algorithm, the attributes of the class are used with a strong assumption that they are conditionally independent of each other. Naïve Bayes provides competitive classification accuracy with this feature. (Webb, 2016).

**k-Nearest Neighbor (KNN):** In some estimates, reliable parametric estimates of probability densities are uncertain or difficult to specify. For this reason, the KNN algorithm was found and discriminant analysis was performed (Peterson, 2009).

**Support Vector Machine (SVM):** Required components for SVM algorithm; the dataset required for training, the class labels and finally the dataset for testing. While making a class label, positive is taken as “1”; negative is taken as “-1” for each sample. (Pavlids, 2004).

**Random Forest (RF):** Random forest is a combination of tree estimators. This combination depends on the values of an individually sampled random vector that has the same distribution for all trees in the forest (Breiman, 2001).

# CHAPTER FIVE
# PROPOSED APPROACH: MULTI-OUTPUT REGRESSSION FOR SOIL MOISTURE PREDICTION (MOR-SMP)

## 5.1 Multi-Output Regression for Soil Moisture Prediction

The aim of the multi-output regression in this study is to estimate soil moisture values at 15, 30, and 45 cm simultaneously based on soil and meteorological input parameters.

Figure 5.1 presents the general structure of the implemented Multi-Output Regression approach in this study. In the first phase, meteorological and soil data is obtained from the sensors. In the next phase, the dataset is passed through a data preprocessing step (one-hot encoding and missing data imputation) to make it ready for the implementation of the proposed method. In the training phase, the Multi-Output Regression approach is applied to the meteorological and soil data. In this step, three different regression models are constructed for predicting soil moisture values at various depths (i.e., 15, 30, and 45 cm), respectively. After that, in the evaluation phase, the soil moisture estimation performing of these models are tested using the *k*-fold cross-validation technique in terms of the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination ($R^2$) metrics. Finally, a new sample is predicted using the Multi-Output Regression method and represented in the presentation step.

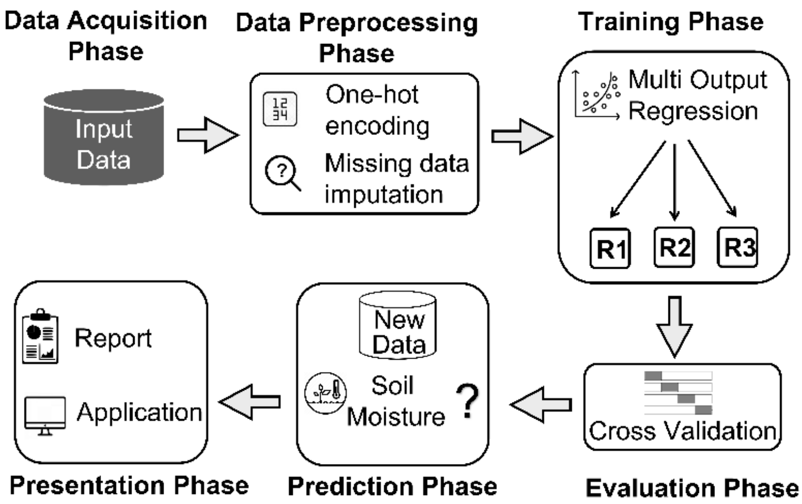


Figure 5.1 The general structure of the Multi-output Regression for soil moisture prediction

## 5.2 Formal Definition

Let $D$ be a meteorological and soil data $D$={$(x_i, y_i)_{i=1}^{N}$ $\epsilon$ $X$ $x$ $Y$} of $N$ instances, where $X$ $\epsilon$ $R^d$ and its related continues-valued multiple outputs are characterized by an output vector $Y$={$(y_i)_{i=1}^{N}$}, where $Y$ $\epsilon$ $R^m$. The aim of the multi-output regression to predict multiple values simultaneously based on given input parameters.

## 5.3 Regression Methods

In this study, nine algorithms (LR, RR, Lasso, RF, ETR, AdaBoost, GB, XGBoost, and HGB) were used when implementing the multi-output regression method.

**Linear Regression (LR):** LR is a statistical regression algorithm that discovers relations between two variables applying a linear equation (Yan & Su, 2009).

**Ridge Regression (RR):** RR is a kind of regularized linear regression that uses L2 penalty for minimizing the size of all coefficients (Vilaming & Groenen, 2015).

**Lasso:** Lasso is a type of linear regression that uses L1 penalty for shrinking the coefficients for input parameters that do not make a great contribution to prediction (Sujatha & Jayanthi, 2020).

**Random Forest (RF):** RF is a bagging algorithm that uses a random feature subset selection method for constructing multiple decision trees (Yildirim et al., 2018).

**Extra Tree Regression (ETR):** ETR constructs multiple unpruned regression trees using the Top-Down approach for the prediction task (Nistane & Harsha, 2018).

**AdaBoost:** AdaBoost trains multiple single split decision trees consecutively to con- vert weak learners to strong ones by reweighting variables in the training set (Yildirim et al., 2019).

**Gradient Boosting (GB):** GB is a boosting algorithm that uses gradients in the loss function to improve the prediction ability of learners in a sequential manner (Biau et al., 2019).

**XGBoost:** XGBoost is an improved distributed gradient boosting method that implements a parallel tree boosting to predict new samples quickly and accurately (Dhaliwal et al., 2018).

**Histogram-based Gradient Boosting (HGB):** HGB is a fast GB model that builds histograms for each feature to get the best split point (Kone et al., 2020).

# CHAPTER SIX
# EXPERIMENTAL STUDIES

## 6.1 Soil Temperature Prediction

In the experimental studies, the proposed STOC method was executed on real-world meteorological and soil data from 16 stations for estimating soil temperature at five soil depths (2, 4, 8, 20, and 40 inches). The proposed application was developed by using the WEKA machine learning library (Witten et al., 2016) in the Java programming language. In this study, five different traditional classification algorithms (DT, NB, KNN, SVM, and RF) were chosen as base learners for the STOC method. Except for the KNN algorithm, all other base learners were implemented with their default parameter values. In each experiment, the $k$ value of the KNN algorithm was selected as $log_2(n)$, where $n$ is the number of instances in each dataset. The proposed STOC models using five different base classifiers were compared by using the $n$-fold cross-validation technique selecting $n$ as 10. The prediction performance of each model was tested in terms of accuracy rate and F-measure metric, and the obtained results were presented via tables and graphs in this study.

Accuracy rate gives a ratio of correctly classified instances to all instances in the dataset as shown in Equation (6.1).

$$Accuracy\ rate = \frac{\#\ of\ correctly\ classified\ instances}{\#\ of\ total\ instances} \tag{6.1}$$

F-measure is a useful metric for presenting the classification performance of any model which is calculated as a harmonic mean of precision and recall values as shown in Equation (6.2).

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \quad (6.2)$$

### *6.1.1 Dataset Description*

In the soil temperature prediction study, 16 different experimental datasets which contain daily values of meteorological and soil data obtained from 16 stations in three states (Utah, Alabama, and New Mexico) of United States at five soil depths (2, 4, 8, 20, and 40 inches) at the interval of 01.01.2011 and 31.05.2020. Figure 6.1 presents the map of the stations in each state of the United States.

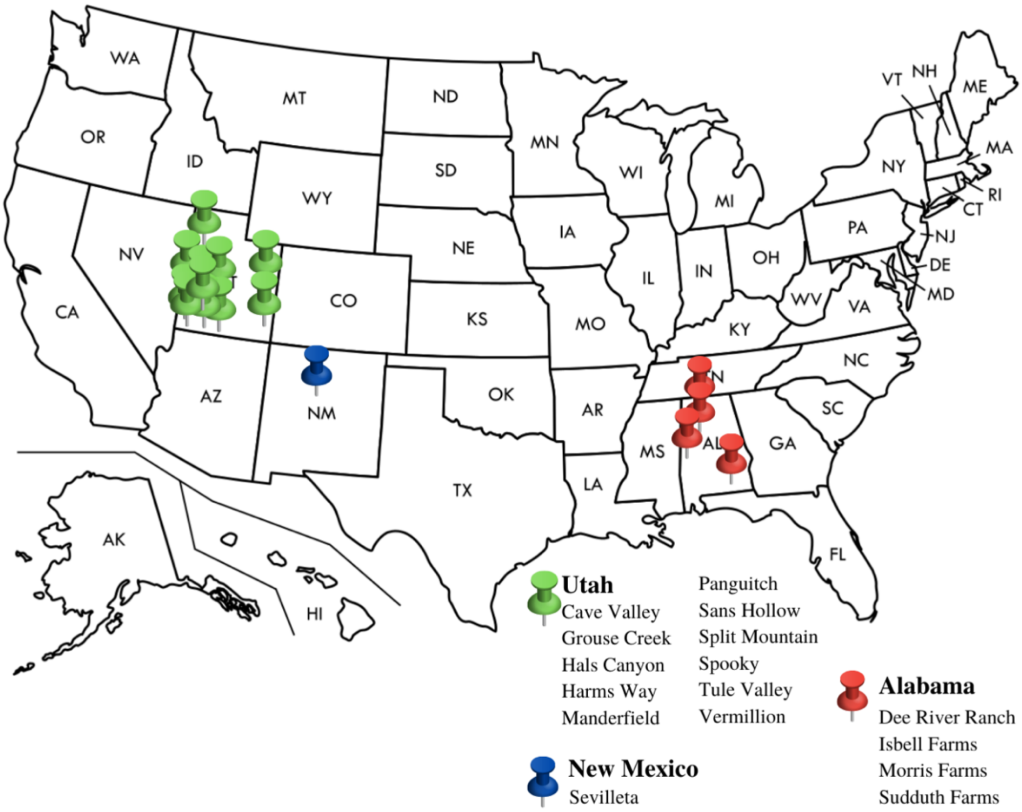


Figure 6.1 The map of the stations used in this study

Each data set in this study includes meteorological parameters such as air temperature, precipitation, relative humidity, solar radiation, wind speed, and vapor pressure, and soil parameters such as moisture and temperature. These real-world and publicly available datasets were gathered from the website (https://www.wcc.nrcs.usda.gov) of the National Water and Climate Center (NWCC) in the National Resources Conservation Service (NRCS) under the United States Department of Agriculture. Table 6.1 presents the details of meteorological and soil parameters for each dataset.

Table 6.1 Parameters of the datasets

| | **Parameters** | **Unit** |
|---|---|---|
| **Meteorological Parameters** | Air Temperature Maximum | degF |
| | Air Temperature Minimum | degF |
| | Precipitation Increment | in |
| | Relative Humidity | pct |
| | Solar Radiation Average | watt/m$^2$ |
| | Solar Radiation/Langley Total | langley |
| | Wind Speed Maximum | mph |
| | Wind Speed Average | mph |
| | Vapor Pressure – Partial | in Hg |
| | Vapor Pressure – Saturated | in Hg |
| **Soil Parameters** | Soil Moisture Percent -2in | pct |
| | Soil Moisture Percent -4in | pct |
| | Soil Moisture Percent -8in | pct |
| | Soil Moisture Percent -20in | pct |
| | Soil Moisture Percent -40in | pct |
| | Soil Temperature Observed -2in | degF |
| | Soil Temperature Observed -4in | degF |

Table 6.1 continues

| | |
|---|---|
| Soil Temperature Observed -8in | degF |
| Soil Temperature Observed -20in | degF |
| Soil Temperature Observed -40in | degF |

The datasets considered in this study have passed through data preprocessing steps using Python Scikit-learn library for the implementation of the STOC method. First, “date” and “station ID” parameters were extracted from each dataset because they are not correlated with the target parameter (soil temperature observed). Then, the hourly collected data were converted into daily data and the instances which have no soil temperature value were eliminated. Furthermore, the proposed STOC method cannot handle continuous target parameters, so, the first, soil temperature observed parameter for each soil depths (2, 4, 8, 20, and 40 inches) in Fahrenheit degree were converted to Celsius degree and then they were discretized into five different categories such as “very low”, “low”, “medium”, “high”, and “very high” according to specific value ranges as shown in Table 6.2.

Table 6.2 Discretization of soil temperature observed parameter

| **Continuous Values (Celsius)** | **Continuous Values (Fahrenheit)** | **Categorical Values** |
|---|---|---|
| **[..., 5]** | [..., 41] | Very low |
| **(5, 10]** | (41, 50] | Low |
| **(10, 15]** | (50, 59] | Medium |
| **(15, 20]** | (59, 68] | High |
| **(20, …]** | (68, …] | Very high |

### *6.1.2 Experimental Results*

In this study, the experiments were carried out to show the classification performance of the STOC approach on real-world meteorological and soil data from 16 stations for estimating soil temperature at five soil depths (2, 4, 8, 20, and 40 inches). The conventional classification algorithms (DT, NB, KNN, SVM, and RF) were selected as base learners for the proposed method separately. To test the success of the proposed STOC method on the datasets, accuracy rate and F-measure values were evaluated.

Table 6.3, 6.4, 6.5, 6.6, and 6.7 present the accuracy rate results of the applied STOC models using the traditional classification algorithms as base learners (STOC.DT, STOC.NB, STOC.KNN, STOC.SVM, and STOC.RF) on the datasets obtained from 16 stations in three states (Utah, Alabama and New Mexico) of United States at five soil depths (2, 4, 8, 20, and 40 inches). The results obtained from these tables indicate that the STOC models using five classification algorithms show different classification performance on these datasets. For example, STOC.KNN and STOC.SVM algorithms achieved the highest accuracy rate value (96.3%) on the dataset with 4 inches depth of the Sudduth Farms and Sevilleta stations, respectively.

Also, when the accuracy rate results take into consideration, it is monitored that STOC.NB demonstrated a worse classification ability than the other STOC models. However, the applied STOC models, except for STOC.NB, show over 75.5% soil temperature prediction performance.

Table 6.3 The accuracy rates (%) of the applied STOC.DT model on 16 stations at five soil depths (inch)

| | **Station ID** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **D** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| 2 | 90.9 | 90.4 | 92.9 | 89 | 92.6 | 92.6 | 88.6 | 95.3 | 95.6 | 94.6 | 94.8 | 93.6 | 93.4 | 92.6 | 92.3 | 92 |
| 4 | 95.3 | 93.2 | 93.5 | 91.1 | 93.4 | 92.1 | 92.6 | 94 | 95.4 | 94.5 | 94.9 | 94 | 94 | 95.2 | 92.6 | 92.7 |

Table 6.3 continues

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8 | 93.9 | 92.9 | 92.4 | 89.1 | 90.4 | 90.1 | 92.9 | 84 | 95.4 | 92.8 | 91.1 | 92.5 | 94.1 | 94.7 | 92 | 91.1 |
| 20 | 87.3 | 89.5 | 88.8 | 88.3 | 88.5 | 84.8 | 88.7 | 89.6 | 91.8 | 88.9 | 87.3 | 91 | 90.3 | 90 | 90.3 | 87.7 |
| 40 | 84.6 | 89.4 | 90.5 | 87.8 | 88.4 | 82 | 90.4 | 84.7 | 93.7 | 86.7 | 88.6 | 87.7 | 89.1 | 85.9 | 85.8 | 89.7 |

Table 6.4 The accuracy rates (%) of the applied STOC.NB model on 16 stations at five soil depths (inch)

| | **Station ID** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **D** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| 2 | 81.1 | 77 | 84.3 | 84 | 83.2 | 77 | 80.3 | 81.8 | 85.3 | 79.9 | 83.3 | 89.4 | 87.1 | 79.3 | 85.6 | 81.4 |
| 4 | 85.1 | 79.7 | 84.5 | 85.3 | 85.5 | 76.3 | 81.5 | 82.8 | 78.5 | 81.7 | 84.9 | 90.7 | 88.3 | 80.2 | 88.1 | 81 |
| 8 | 85.9 | 79.5 | 82.5 | 83.2 | 82.6 | 81.2 | 80.8 | 74.4 | 89.1 | 80.8 | 80.8 | 88.1 | 87.5 | 79.3 | 85.9 | 80.9 |
| 20 | 75.4 | 78 | 72.6 | 71.2 | 75.3 | 73.4 | 68.4 | 80.8 | 77.3 | 73.8 | 75 | 76.3 | 75.4 | 76.5 | 78.6 | 66.7 |
| 40 | 63.5 | 75.9 | 59 | 55.7 | 51.7 | 63.3 | 65.6 | 72.6 | 69.5 | 70.5 | 58.6 | 58.3 | 64 | 67.1 | 60.5 | 71 |

Table 6.5 The accuracy rates (%) of the applied STOC.KNN model on 16 stations at five soil depths (inch)

| | **Station ID** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **D** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| 2 | 91.9 | 91.2 | 92 | 88.6 | 93.2 | 92.1 | 88.8 | 95.1 | 95.7 | 94.3 | 94.9 | 93.4 | 93.6 | 93.4 | 91.2 | 92.7 |
| 4 | 95 | 94 | 94.7 | 90.9 | 93.7 | 93.2 | 92.4 | 95 | 95.7 | 95.5 | 96 | 93.1 | 94.9 | 96.3 | 92.8 | 93.4 |
| 8 | 95 | 93.7 | 92.9 | 89.6 | 91.5 | 90.3 | 93.5 | 83.4 | 95.5 | 93.8 | 91.4 | 93.6 | 94.5 | 95.4 | 92.2 | 92.2 |
| 20 | 87.4 | 88.7 | 88.3 | 86.8 | 87 | 87.3 | 86.7 | 90.4 | 88.1 | 88.2 | 87.4 | 89.6 | 89 | 89.7 | 89.9 | 86.3 |
| 40 | 83.5 | 87.3 | 86.9 | 82 | 82.9 | 82.4 | 84.5 | 85.8 | 89.6 | 85.9 | 85.8 | 84.3 | 85.7 | 86.1 | 81.6 | 86.4 |

Table 6.6 The accuracy rates (%) of the applied STOC.SVM model on 16 stations at five soil depths (inch)

| | Station ID | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **D** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| 2 | 92.1 | 90.4 | 93.1 | 90.2 | 93.6 | 92.3 | 90.4 | 95.4 | 95.8 | 94.9 | 95.6 | 94.1 | 93.9 | 93.5 | 92.2 | 93.3 |
| 4 | 95.2 | 93.9 | 94.8 | 92.6 | 94.5 | 92.6 | 92.9 | 95.2 | 95.9 | 95.4 | 96.3 | 93.9 | 95 | 95.9 | 93.6 | 93.8 |
| 8 | 95.3 | 93.8 | 93 | 89.8 | 91.5 | 89.3 | 93.2 | 83.4 | 95.6 | 94 | 91.3 | 94.1 | 94.8 | 95.6 | 93.4 | 92.3 |
| 20 | 86.6 | 85 | 87.7 | 86.5 | 87.4 | 86.1 | 84.9 | 89.7 | 87.8 | 87.9 | 85.3 | 90.1 | 89.5 | 88.7 | 90.4 | 86.3 |
| 40 | 80.2 | 85.7 | 85.3 | 75.5 | 83.1 | 80.8 | 83.4 | 83.9 | 81.8 | 83.9 | 80.8 | 79 | 81.8 | 85.5 | 77 | 84.7 |

Table 6.7 The accuracy rates (%) of the applied STOC.RF model on 16 stations at five soil depths (inch)

| | Station ID | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **D** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| 2 | 92.5 | 92.2 | 93.2 | 90.2 | 93.2 | 93.6 | 89.7 | 95.4 | 96 | 95.3 | 95.7 | 94.1 | 94.3 | 93.5 | 92.8 | 92.9 |
| 4 | 95.5 | 94.1 | 94.1 | 91.8 | 94 | 92.6 | 92.6 | 95.3 | 95.8 | 95.4 | 95.7 | 94.4 | 95 | 96.1 | 93.5 | 93.2 |
| 8 | 94.8 | 94 | 92.2 | 90.3 | 90.8 | 89.5 | 93.6 | 84.5 | 95.2 | 93.7 | 91.8 | 93.6 | 94.6 | 95.9 | 93.3 | 92.1 |
| 20 | 87.6 | 87 | 87.6 | 88.7 | 87.7 | 87 | 85.4 | 90.7 | 87.7 | 88.4 | 87.2 | 89.5 | 89.7 | 90.5 | 89.7 | 84.9 |
| 40 | 83.3 | 87.9 | 86 | 85.7 | 85.4 | 81.1 | 86.9 | 84.2 | 87.7 | 85.9 | 86.5 | 82.4 | 86.5 | 85.1 | 83.6 | 85.5 |

The average accuracy rates of the STOC models on these datasets were calculated to reveal which model is more successful in soil temperature estimation. The obtained values are illustrated in the graph given in Figure 6.2. The results from this graph indicate that STOC.DT achieved the best soil temperature prediction performance with an accuracy rate of 90.95%. Also, it is possible to say that STOC.RF and STOC.KNN reached higher classification abilities with quite close accuracy rates of 90.91% and 90.84%, respectively.

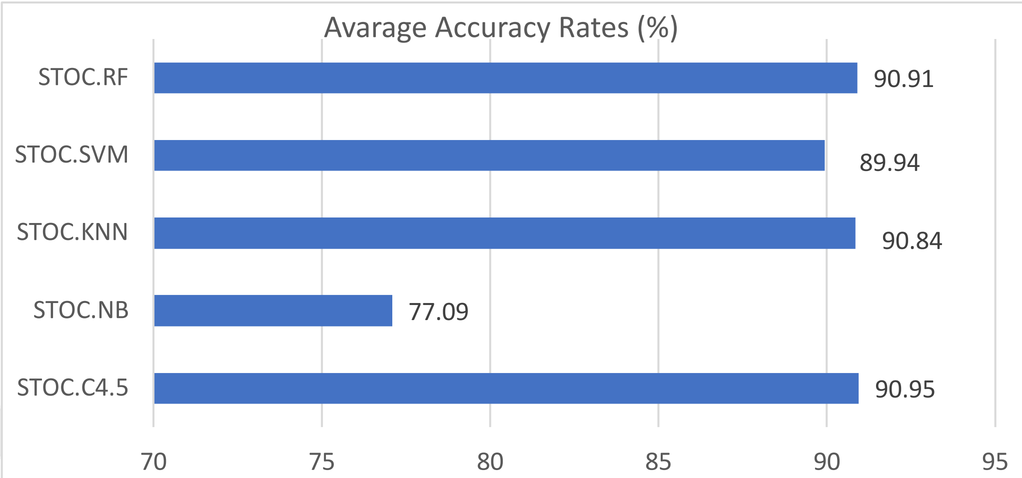


Figure 6.2 The average accuracy rates (%) of the applied STOC models on the 16 different stations

Furthermore, the F-measure values of the applied STOC models using the traditional classification algorithms on the 16 datasets obtained in three of United States at five soil depths (2, 4, 8, 20, and 40 inches) were evaluated. In Figure 6.3, 6.4, and 6.5, the average F-measure values of the STOC models for each state (Utah, Alabama and New Mexico) were illustrated, respectively. The fact that the F-measure value is close to 1 presents that the model offers a successful classification ability. These graphs in the figures indicated that the applied all STOC models, except for STOC.NB, provide over 0.8 F-measure value. Therefore, it is possible to say that the proposed STOC approach is a successful method for soil temperature prediction of ordinal data.

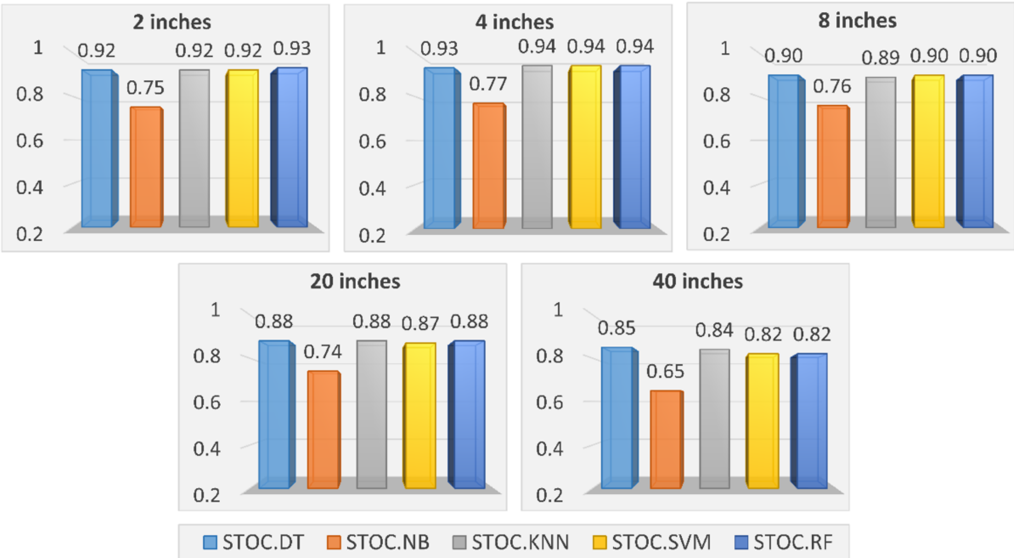


Figure 6.3 The average F-measure values of the applied STOC models on the Alabama state stations

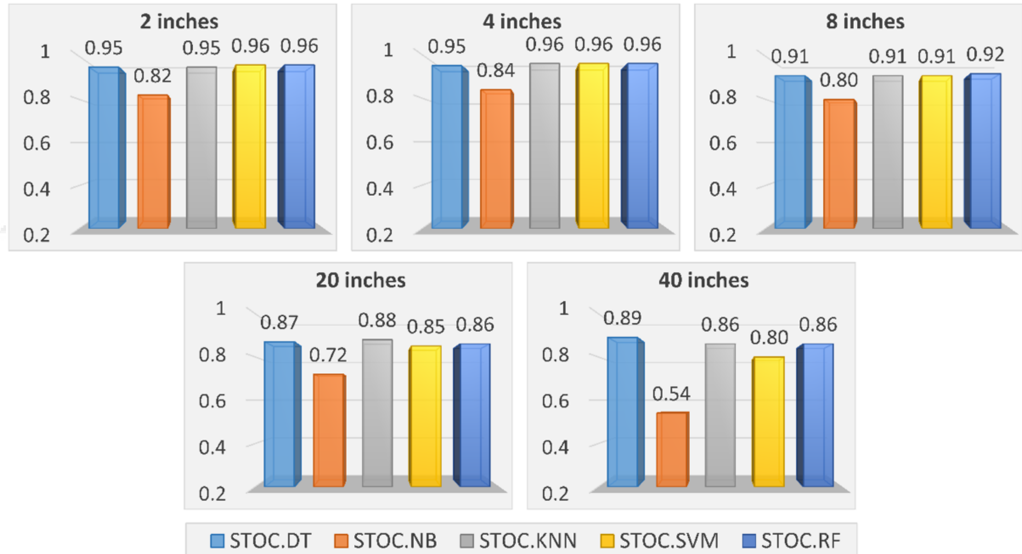


Figure 6.4 The average F-measure values of the applied STOC models on the Mexico state stations

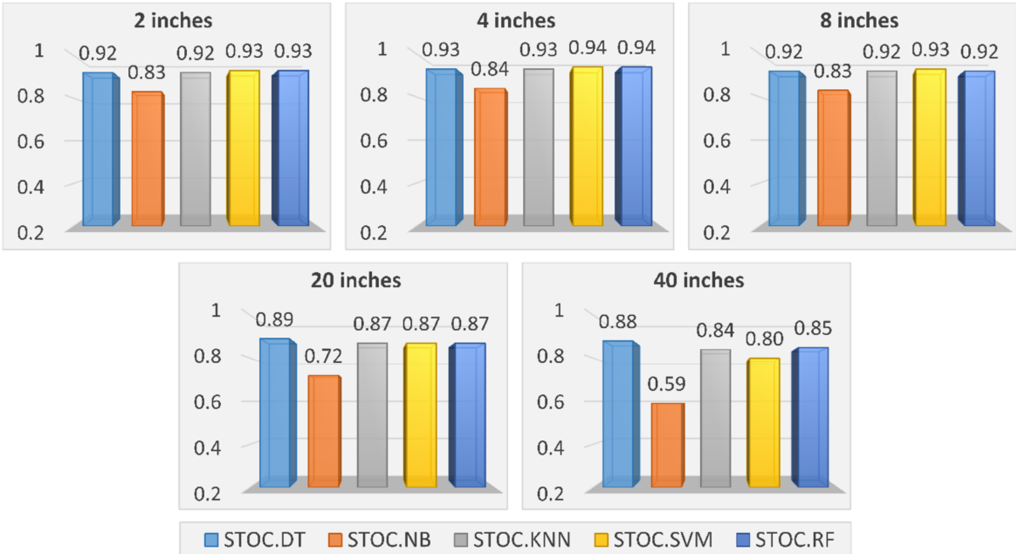


Figure 6.5 The average F-measure values of the applied STOC models on the Utah state stations

### 6.2 Soil Moisture Prediction

In the experimental studies, the multi-output regression models were constructed on real-world meteorological and soil data for estimating soil moisture at three soil depths (5, 30, and 45 cm). The soil prediction application was developed in Python using Scikit-learn, Numpy, and Pandas libraries. In this study, the multi-output regression models were generated using the LR, RR, Lasso, RF, ETR, AdaBoost, GB, XGBoost, and HGB algorithms. The applied algorithms' parameters were set as follows:

- **LR:** positive=True
- **RR:** random_state=1
- **Lasso:** random_state=1, alpha=0.1
- **ETR:** random_state=1
- **RF:** random_state=1, criterion=MAE
- **AdaBoost:** random_state=1, n_estimators=10
- **GB:** random_state=1, n_estimators=1000, criterion=MSE, max_feature=auto
- **HGB:** random_state=1, max_bins=50

All other parameters of the applied algorithms were chosen as default values. The multi-output regression models were compared by using the *k*-fold cross-validation technique by selecting $k$ as 10. The validation process was repeated three times, and the average of them were taken as a validation result. The capabilities of these models were tested using the MAE, RMSE, and $R^2$ measures.

**Mean Absolute Error (MAE):** It gives the average magnitude of difference between the actual and predicted values as shown in Equation (1).

$$\frac{1}{m}\sum_{i=1}^{m}|(a_i - p_i)| \tag{6.3}$$

**Root Mean Squared Error (RMSE):** It gives the square root of the mean of square of difference between the actual and predicted values as shown in Equation (2).

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(a_i - p_i)^2} \tag{6.4}$$

**Coefficient of Determination ($R^2$):** It gives a degree ranging from 0 to 1.0 for presenting how strong the linear relationship is between the actual and predicted values, as shown in Equation (3).

$$R^2 = \frac{\left(\frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{m-1}\right)^2}{\left(\sqrt{\frac{\sum_i (p_i - \bar{p})^2}{m-1}\frac{\sum_i (a_i - \bar{a})^2}{m-1}}\right)^2} \tag{6.5}$$

where $p$ is predicted target value, $a$ represents actual target value, and $m$ is the number of instances.

### *6.2.1 Dataset Description*

In this study, a real-world dataset that consists of daily values of meteorological and soil data obtained from Kemalpaşa-Örnekköy station in Turkey at three soil depths (15, 30, and 45 cm) at the interval of 01.01.2017 and 28.12.2019. Figure 6.6 presents the map of the station in Izmir, Turkey.


Figure 6.6 The map of the Kemalpaşa-Örnekköy station used in this study

The dataset used in this study includes the date, meteorological features such as air temperature, wet bulb temperature, dew point temperature, solar radiation, vapor pressure deficit, relative humidity, precipitation, leaf wetness, wind speed, and soil features such as temperature and moisture. It has passed through data preprocessing steps using the Python Scikit-learn library to implement the proposed method. First, "month" and "season" features were extracted from the "date" variable in the dataset. Then, these newly added features were converted to numerical values using the one-hot encoder technique. Furthermore, all missing values of the features were imputed using mean values of the same feature's existing values.

### *6.2.2 Experimental Results*

Table 6.8 presents the MAE and RMSE values obtained from the Multi-Output Regression models with their standard deviation (Std.) values. According to these results, ETR achieved the best prediction performance among the others with the values of 22.70 MAE and 37.35 RMSE. When these results are considered in general, it is understood that ensemble learning algorithms such as RF, ETR, AdaBoost, GB, XGBoost, and HGB outperform the traditional individual algorithms.

Table 6.8 Comparison of MSE and RMSE values of the implemented models

| | | **Individual Algorithms** | | | **Ensemble Algorithms** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Metrics** | | **LR** | **RR** | **Lasso** | **RF** | **ETR** | **Ada Boost** | **GB** | **XG Boost** | **HGB** |
| **MAE** | Value | 37.09 | 37.43 | 37.39 | 25.15 | **22.70** | 36.62 | 27.25 | 27.83 | 25.56 |
| | Std. | 2.97 | 3.06 | 3.13 | 2.74 | 2.92 | 2.41 | 3.05 | 2.82 | 2.66 |
| **RMSE** | Value | 50.08 | 50.61 | 50.65 | 39.96 | **37.35** | 47.44 | 41.35 | 41.58 | 39.03 |
| | Std. | 3.98 | 5.51 | 5.62 | 4.26 | 4.33 | 2.87 | 4.78 | 4.32 | 4.09 |

The graph given in Figure 6.7 illustrates the $R^2$ values of the Multi-output regression methods on the experimental dataset. If the $R^2$ value of any algorithm is close to 1, it means that this model is highly reliable for the prediction task. The results indicate that the ETR model achieved the best soil moisture prediction performance with an $R^2$ value of 0.81. In addition, it is observed from this graph that all algorithms applied in this study showed over 0.65 $R^2$ values. Thus, it can be concluded that the Multi-Output Regression method can successfully predict soil moisture at different depths.

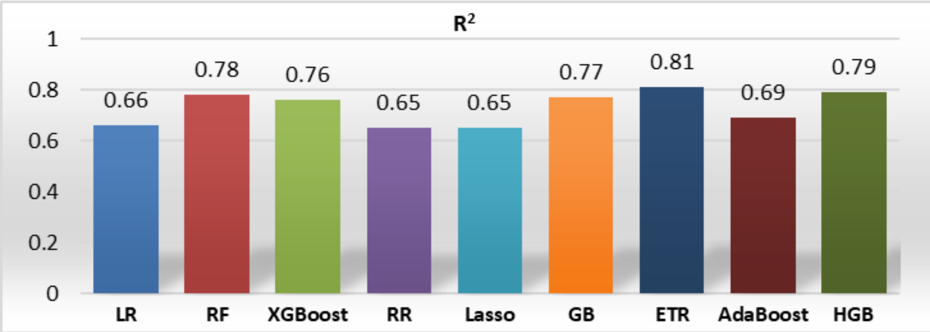


Figure 6.7 Comparison of $R^2$ values of the implemented models

## CHAPTER SEVEN
## CONCLUSION AND FUTURE WORKS

### 7.1 Conclusion

Recently, with the developments in technology, there have been important changes in this sense in the agricultural sector. In particular, technologies such as the Internet of Things (IoT) play a major role in shaping the future of agriculture. In the smart farming systems, it is aimed to transfer and store the data received from the sensors over the internet, and to obtain more crops with less cost within the information obtained from these data. Two of the most important factors in obtaining a crop are soil temperature and soil moisture. *Soil temperature* is important for the germination of the seed, for the plant to carry out its vital activities, and not to adversely affect the chemical and physical properties of the plant. *Soil moisture* is important because the amount of water that the plant needs to take should be kept at a certain rate. Therefore, in this thesis, two separate studies have been conducted on soil temperature and soil moisture estimation using machine learning algorithms.

*Soil temperature prediction* plays an important role in the agriculture sector. In the literature, some machine learning-based studies performed regression or nominal classification tasks. However, the target attribute values have a natural order and considering the order in the class values can affect the soil temperature estimation performance. Considering this motivation, a novel soil temperature ordinal classification (STOC) approach is proposed in this study, which categorizes the soil temperature levels of meteorological and soil data at different soil depths by taking into account the inherent order of class labels. The proposed approach was tested with different base learners, including DT, NB, KNN, SVM, and RF. In the experiments, the proposed STOC method with different classifiers (STOC.DT, STOC.NB, STOC.KNN, STOC.SVM, and STOC.RF) was applied on daily values of meteorological and soil data obtained from 16 stations at five soil depths (2, 4, 8, 20, and 40 inches) and compared with each other in terms of F-measure and accuracy rate metrics. The results show that the proposed STOC method provides accurate soil temperature classification performance. The STOC models implemented in this study, except for STOC.NB, achieved over 75.5% soil temperature estimation ability. Also, it is clearly understood from the experimental result that the STOC.DT shows the best soil temperature prediction performance with an average accuracy rate of 90.95%.

*Soil moisture prediction* has a significant impact on especially agricultural production and water recourses management. Because of this reason, this study implements a Multi-Output Regression method on daily values of meteorological and soil data for predicting soil moisture at 15, 30, and 45 cm depths. In this study, this approach was tested by using the LR, RR, Lasso, RF, ETR, AdaBoost, GB, XGBoost, and HGB algorithms. The Multi-Output Regression (MOR) models were compared each other in terms of MAE, RMSE, and $R^2$ in the experiments. The results presented that the MOR approach gave successful prediction performance on soil moisture. Also, the ETR algorithm showed the most successful result among all other algorithms.

### 7.2 Future Works

As a future work, an improved time-series technique can be used to solve the soil temperature prediction and soil moisture prediction problems. In addition, a multi-output regression technique can be implemented on the same meteorological and soil data for predicting continuous soil temperature values. Furthermore, an ensemble of Multi-Output Regression could also be applied and compared with the proposed model in the future. On the other hand, using the libraries that support machine learning algorithms in Apache Spark, which is a lightning-fast cluster computing, a study can be carried out on the instant classification of big data.

# REFERENCES


Abyaneh, H. Z., Varkeshi, M. B., Golmohammadi, G. & Mohammadi, K. (2016). Soil temperature estimation using an artificial neural network and co-active neuro-fuzzy inference system in two different climates. *Arabian Journal of Geosciences, 9*(2016), 1-9.

Alizamir, M., Kim, S., Kermani, M. Z., Heddam, S., Shahrabadi, A. H. & Gharabaghi, B. (2020). Modelling daily soil temperature by hydro-meteorological data at different depths using a novel data-intelligence model: deep echo state network model. *Artificial Intelligence Review, 102*(2020), 1-28.

Alizamir, M., Kisi, O., Ahmed, A. N., Mert, C., Fai, C. M., Kim, S., Kim, N. W. & Shafie, A. E. (2020). Advanced machine learning model for better prediction accuracy of soil temperature at different depths. *Plos One, 15*(4), 1-25.

Breiman, L. (2001). Random forests (machine learning), *2001 Kluwer Academic Publishers, Manufactured in The Netherlands, 45*(2001), 5-32.

Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., Xue, X.(2019). Research on soil moisture prediction model based on deep learning. *Plos One, 14*(2019), e0214508.

Feng, Y., Cui, N., Hao, W., Gao, L. & Gong, D. (2019). Estimation of soil temperature from meteorological data using different machine learning models. *Geoderma, 338*(2019), 67-77.

Frank, E. & Hall, M. (2001). A simple approach to ordinal classification. *In: European Conference on Machine Learning 3-5 September, Freiburg, Germany*, 145-156.

Hong, Z., Kalbarczyk, Z. & Iyer, R. K. (2016). A data-driven approach to soil moisture collection and prediction. *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, 1-6.

Karakaş, M. (2020). *İşgücü İstatistikleri.* Retrieved March 31, 2021, from https://data.tuik.gov.tr/Bulten/Index?p=Isgucu-Istatistikleri-Ocak-2020-33785

Kisi, O., Tombul, M. & Kermani, M. Z. (2015). Modeling soil temperatures at different depths by using three different neural computing techniques. *Theoretical and Applied Climatology*, *121*(2015), 377-387.

Kucuk, C., Birant, D. & Taser, Y. P. (2021). A multi-output regression model for soil moisture prediction. *International Conference on Intelligent and Fuzzy Systems* & *INFUS2021, 304*(2021), 1-2.

Li, C., Zhang, Y. & Ren, X. (2020). Modeling hourly soil temperature using deep BiLSTM neural network. *Algorithms*, *13*(7), 173-187.

Liu, G., Lin, Z. & Yu, Y. (2009). Multi-output regression on the output manifold. Shanghai Jiao Tong University, Shanghai 200240, PR China, *Pattern Recognition 42*, 2737-2743.

Matei, O., Rusu, T., Petrovan, A. & Mihut, G. (2017). A data mining system for real time soil moisture prediction. *10th International Conference Interdisciplinary in Engineering,* (2017), 837-844.

Mehdizadeh, S., Fathian, F., Safari, M. J. & Khosravi, A. (2020). Developing novel hybrid models for estimation of daily soil temperature at various depths. *Soil & Tillage Research, 197*(2020), 1-12.

Nanda, A., Sen, S., Sharma, A. N. & Sudheer, K. P. (2020). Soil temperature dynamics at Hillslope scale-field observation and machine learning-based approach. *Water, 12*(2020), 713-734.

Onwuka, B. & Mang, B. (2018*).* Effects of soil temperature on some soil properties and plant growth. *Advances in Plants & Agriculture Research 8*(1), 34-37.

Pavlids, P., Wapinski, I. & Noble, W. S. (2004). Support vector machine classification on the web. *Bioinformatics Applications Note, 20/4*(2004), 586–587.

Petersen, L. E (2009). K-nearest neighbor. *Scholarpedia, 4*(2), 1883.

Prasad, R., Deo, R. C., Li, Y. & Mareseni T. (2018). Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. *Geoderma, 330* (2018), 136-161.

Sanikhani, H., Deo, R. C., Yaseen, Z. M., Eray, O. & Kisi, O. (2018). Non-tuned data intelligent model for soil temperature estimation: A new approach. *Geoderma, 330*(2016), 52-64.

Sanuade, O., Hassan, A., Akanji, A., Olaojo, A., Oladunjoye, M., Abdulraheem, A. (2020). New empirical equation to estimate the soil moisture content based on thermal properties using machine learning techniques. *Arabian Journal of Geosciences*, *13*, 377.

Sattari, M. T., Avram, A., Apaydin, H. & Matei, O. (2020). Soil temperature estimation with meteorological parameters by using tree-based hybrid data mining models. *Mathematics 8*(9), 1-21.

Shamshirband, S., Esmaeilbeiki, F., Zarehaghi, D., Neyshabouri, M., Samadianfard, S., Ghorbani, M. A., Mosavi, A., Nabipour, N. & Chau, K. W. (2020). Comparative analysis of hybrid models of firefly optimization algorithm with support vector machines and multilayer perceptron for predicting soil temperature at different depths. *Engineering Applications of Computational Fluid Mechanics 14*(1), 939-953.

Singh, G., Sharma, D., Goap, A., Sehgal, S., Shukla, A., Kumar, S. (2019). Machine Learning based soil moisture prediction for Internet of Things based Smart Irrigation System. *IEEE International Conference on Signal Processing, Computing and Control* (ISPCC 2k19), 175-180.

Tarım ve Orman Bakanlığı (2020), *Tarım Alanları 2001-2020,* Retrieved March 31, 2021, from https://tuikweb.tuik.gov.tr/UstMenu.do?metod=temelist

Webb, G. I. (2016), *Naïve bayes*. Australia: Encyclopedia of Machine Learning and Data Mining, Springer.

Yadav, S. K., & Pal, S. (2012). Data mining: a prediction for performance improvement of engineering students using classification, *World of Science and Information Technology Journal (WCSIT),* 2(2), 51-56.

Yamaç, S., Şeker, C., Negiş, H.(2020). Evaluation of machine learning methods to predict soil moisture constants with different combinations of soil input data for calcareous soils in a semi-arid area. *Agricultural Water Management, 234*(2020), 106121.

Zeynoddin, M., Ebtehaj, I. & Bonakdari, H. (2020). Development of a linear based stochastic model for daily soil temperature prediction: One step forward to sustainable agriculture. *Computers and Electronics in Agriculture 176*(2020), 1-24.

Zhen, X., Wang, Z., Yu, M. & Li, S. (2015). Supervised descriptor learning for multi-output regression. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1211-1218.