Predicting soil moisture based on the color of the leaves using data mining and machine learning techniques

S S Atanasov

1 Trakia University, Faculty of Technics and Technologies, 38 Graf Ignatiev Str., Yambol, Bulgaria

e-mail: svetoslav.atanassov@trakia.uni.bg

Abstract. This research article’s aim is by using data mining and finding a suitable machine learning algorithm (MLA) to predict soil moisture, therefore the need for watering. Prediction is based on a training data set (including color RGB values taken from the leaves and values for soil moisture and soil temperature). A classifier is trained first, on its base a model is created and stored. Finally, with a different test data set, the efficiency of the selected model is checked. The object of study is the color of leaves of indeterminate greenhouse tomato plants of the Panekra variety. According to preliminary assumptions, the most informative about the need for watering are the young leaves (on top of the plant). Among the wide variety of data mining tools, we chose Weka Workbench. The last task of this study is to compare received with the methods of machine learning model and the model obtained in a previous study. For greater completeness of this research, the training of the classifier has been performed both with the whole training data set and with smaller data sets filtered by certain criteria (young/old leaves, etc.). The ultimate goal is water-saving and optimizing watering and water using. The resulting model is efficient and predicts based on the color of the young leaves with 0-5% error, and by 8-12%, based on the color of the old ones, before watering, taking into account the influence of soil temperature into the model.

1. Introduction

Water-saving and its reasonable and optimal use for irrigation purposes are becoming in past years the main goal of many scientific types of research and implementations worldwide, as well as this one. Vegetables are the thirstiest in the fruit-growing and absorb large amounts of water during their growing season. Physiological studies have found, that throughout their period of growing tomatoes need hundreds of litres water per square meter. The daily water consumption is very high. Tomatoes need moderate soil moisture because if you overdo it with watering they quickly form a powerful vegetative mass and fewer fruits. Over-wetting the soil is the other extreme, which should not be allowed. Vegetables are very sensitive to the lack of air in the soil.

Conventional methods for managing the irrigation process are: sensing the soil water status [1] and sensing the plant water status [2]. Direct and indirect methods of soil water status determination [3, 5, 6] include the gravimetric method, neutron probe, dielectric methods (TDR, FDR), tensiometers, resistance blocks, thermal heat probe, and soil psychrometer. Methods of sensing the plant water status [7-9] are pressure chamber, ZIM-probe, dendrometers, heat balance/pulse methods, the infrared thermometry measures, porometers, and the infrared gas analyzers. Last but not least, there is probably the oldest and simplest method for determining soil moisture - by the subjective perception by squeezing and pressing soil sample by hand.
Based on the accumulated data from direct and indirect measurements of soil water status and plant water status it is possible to create predictive mathematical models to be used in determining the need for watering: using classical statistical methods [10], used these days machine learning techniques [11-16] with help of different classification models (for example random forest, neural network, support vector machine, etc.), combined - using classic methods and machine learning [17, 18]. Data mining is the process of finding patterns and correlations within data sets, used to predict numeric values or to classify nominal values.

Regarding the use of leaves in predicting the water status of plants and the implementation of precise irrigation there are scientific researches [19, 20] using leaf thickness and electrical capacitance as measures of plant water status. Others [21] continuously monitor canopy temperature as an indicator of plant water status. Another one [22] offers precision irrigation in wine grape using a proximal leaf monitor system for measuring plant water status.

In general, there is a lack of information the color of the leaf to be used as an indicator of the need for irrigation or any water stress. Despite the concerns of some scientists [1, 2] that the color of the leaves is a late indicator of water stress we, based on additional observations on the time of absorption of water and nutrients from tomatoes [23], have found that the tomato plant reacts to the change of moisture within a maximum of 2 hours (the water travels 1 meter in 1 hour). So if a continuous monitoring system is developed the color of the leaves will not be a late indicator. Of course, this does not include cases of infected or diseased plants with changed color of the leaves.

Continuous monitoring systems [21, 24, 25] for constant monitoring and systems for advising the farmer on the need for irrigation can and are built based on the created predictive models. With their help, concepts such as smart farming [26], precision irrigation [22, 25], speaking plant [27], widely implementing of WSN and IoT in process of irrigation [39] arise.

This article offers a little more different approach, based on physical observation of plants. Our task is to offer an indirect intelligent, fast method, confirming the long-term observations of farmers.

In general, innovations in this area are a source of competitive advantages and are a driving force for business [28].

2. Materials and methods

The object of the study is the leaf mass of tomato plants (Solanum Lycopersicum) (figure 1) and its influence on the microclimate parameters soil moisture and temperature in the greenhouse.

The studies were performed in a certified tomato greenhouse located in the city of Plovdiv (Bulgaria), where all environmental parameters and the health of plants are kept in an optimal range. Studied tomato plants are from one indeterminate variety – named “Panekra”, planted in August. Color measurements are conducted in the second half of September in different meteorological and climatic conditions - in hot and sunny weather, in cloudy, rainy and cooler weather, 24 hours before watering and 24 hours after irrigation.

Figure 1. General view of the greenhouse and the test plants.
The used equipment (spectral colorimeter RGB-1002 and an instrument for measuring environmental parameters PCE-EM 883), as well as the preliminary modeling with another nonlinear method compared later with this study, are described in [10]. The soil moisture sensor and its calibration for the specific soil type are described in [4].

During each measurement from the top of the plant with a colorimeter are taken randomly six RGB values, and from the bottom part - three. This is because based on preliminary assumptions the young leaves on top of the plant should be the best carriers of information about the need for irrigation (that's why we took twice as many RGB color samples from them). Based on the same observations when dried, the tomato leaves darken and are not bright green. The color values are taken at random from different leaves at the individual measurements. At the beginning of the research period, the plants were 150 cm tall and at the end of the measurement period - 190 cm.

The soil in the greenhouse is an alluvial meadow. Its main water-physical property in the layer 0–40 cm, which interests us most in terms of the need for watering, is moisture on FC which is 30.9% volumetric water content [29]. The alluvial meadow soil has a low usable water reserve (TAW=116 mm m⁻³). It consists mainly of sand – between 68% of the surface and 53% and the lower-lying horizons. The content of clay is between 9% and 22%. [30].

The task is with the means of data mining to find a suitable machine learning algorithm for soil moisture prediction based on a training data set (comprising color RGB values and data for soil moisture and temperature - Table 1), to train a classifier, to create and store an appropriate model and to test the model with different testing data set. Then to compare the result with the results obtained in another statistical predictive modeling [10].

The following figure is a table with sample input data, the whole training set is over 200 measurements. Soil temperature is represented in °C, and soil moisture – in % volumetric water content.

**Table 1. Sample of training data set.**

| №  | R   | G   | B   | Irrig | Leaf | Soil Temp | Soil Mois | №  | R   | G   | B   | Irrig | Leaf | Soil Temp | Soil Mois |
|----|-----|-----|-----|-------|------|-----------|-----------|----|-----|-----|-----|-------|------|-----------|-----------|
| 1  | 123 | 143 | 91  | after | young | 20,89     | 27,28     | 22 | 161 | 182 | 134 | before | old  | 19,44     | 22,55     |
| 2  | 148 | 167 | 126 | after | old   | 20,89     | 27,28     | 23 | 101 | 125 | 74  | before | young | 18,51     | 24,64     |
| 3  | 122 | 144 | 80  | after | young | 20,94     | 24,31     | 24 | 148 | 172 | 117 | before | old   | 18,51     | 24,64     |
| 4  | 130 | 149 | 101 | after | old   | 20,94     | 24,31     | 25 | 87  | 106 | 60  | before | young | 19,03     | 20,02     |
| 5  | 121 | 143 | 93  | after | young | 20,94     | 27,83     | 26 | 121 | 139 | 90  | before | old   | 19,03     | 20,02     |
| 6  | 129 | 152 | 103 | after | old   | 20,94     | 27,83     | 27 | 129 | 155 | 93  | before | young | 19,13     | 24,31     |
| 7  | 102 | 122 | 79  | after | young | 20,72     | 25,52     | 28 | 126 | 143 | 105 | before | old   | 19,13     | 24,31     |
| 8  | 145 | 170 | 102 | after | old   | 20,72     | 25,52     | 29 | 131 | 158 | 92  | before | young | 18,70     | 26,73     |
| 9  | 111 | 130 | 77  | after | young | 21,10     | 24,53     | 30 | 129 | 153 | 99  | before | old   | 18,70     | 26,73     |
| 10 | 143 | 171 | 99  | after | old   | 21,10     | 24,53     | 31 | 110 | 133 | 81  | before | young | 19,22     | 21,67     |
| 11 | 131 | 155 | 90  | after | young | 21,65     | 28,16     | 32 | 131 | 150 | 104 | before | old   | 19,22     | 21,67     |
| 12 | 143 | 163 | 97  | after | old   | 21,65     | 28,16     | 33 | 129 | 155 | 96  | before | young | 19,03     | 24,20     |
| 13 | 104 | 124 | 94  | after | young | 21,27     | 26,62     | 34 | 137 | 153 | 118 | before | old   | 19,03     | 24,20     |
| 14 | 137 | 155 | 106 | after | old   | 21,27     | 26,62     | 35 | 90  | 108 | 68  | before | young | 20,72     | 25,74     |
| 15 | 141 | 165 | 95  | after | young | 21,37     | 23,87     | 36 | 138 | 159 | 111 | before | old   | 20,72     | 25,74     |
| 16 | 166 | 187 | 133 | after | old   | 21,37     | 23,87     | 37 | 114 | 136 | 90  | before | young | 19,03     | 22,00     |
| 17 | 116 | 137 | 89  | after | young | 21,80     | 31,02     | 38 | 118 | 135 | 90  | before | old   | 19,03     | 22,00     |
| 18 | 149 | 174 | 102 | after | old   | 21,80     | 31,02     | 39 | 106 | 130 | 74  | before | young | 18,96     | 23,21     |
| 19 | 116 | 146 | 85  | before | young | 19,18     | 24,75     | 40 | 162 | 183 | 135 | before | old   | 18,96     | 23,21     |
| 20 | 159 | 181 | 132 | before | old   | 19,18     | 24,75     | 41 | 105 | 126 | 76  | before | young | 18,89     | 21,45     |

3
Almost all known software applications for statistical processing (Statsoft Statistica, SPSS, SAS) contain data mining tools powered by machine learning algorithms. Data mining may be conducted also in programming environments such as R and Python, as well as in software environments such as Machine Learning Server Windows and Matlab. There are also available applications specialized in machine self-learning and data mining such as Weka, Rapidminer, Tanagra, and Orange.

The Weka workbench is an open-source machine learning software under GNU license that includes methods for the main data mining problems: regression, classification, clustering, association rule mining, and attribute selection. We will use Weka because the following reasons and benefits that we believe it possesses [31]:

- 100+ algorithms for classification,
- 75 for data pre-processing,
- 25 to assist with feature selection,
- 20 for clustering, finding association rules, etc.

Last but not least, Weka offers many options for additional customizing and adjusting of the classifiers, well noted and wonderfully explained in the help.

3. Results
In the form of a sequence of steps is described procedure for training classifiers with the training data set using regression algorithms. Our data set consists of numerical data and we predict numerical values, i.e. we have a regression problem. For greater completeness and comprehensiveness of the study, the training of the classifier was performed both with the whole training data set (named in the tables full set), as well with smaller datasets filtered by certain criteria (young leaves/old leaves, 24 hours before/24 hours after watering).

3.1. Normalizing and standardizing
In [10] a check for normal (Gaussian) distribution is made and it is present in most color components of the RGB color model. Only the color components B for young and old leaves after watering and G in young leaves before watering do not have Gaussian distribution.

If Gaussian distribution is missing, Weka provides an opportunity to normalize the data with numerical values [32]. Data normalization is the process of rescaling of attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.

For nominal data, Weka offers data standardization. Data standardization is the process of rescaling one or more attributes so that they have a mean value of 0 and a standard deviation of 1. In standardization, it is assumed that there is already a normal distribution.

There is also an option numeric attributes to be discretized - i.e. to be made nominal.

If we only have two nominal values, they can be converted into numerical - 0 and 1.

For the above purposes are used unsupervised filters – they can be applied without user control.

Weka also offers four options for the input data set testing [33-35]. In our study we will use the "Cross-validation" option in which by default the set is divided into 10 parts, the selected classifier is trained with 9 of them and the model is tested with the last one. The classifier is evaluated by cross-validation, using the number of folds that are entered in the Folds text field. With 10-fold cross-validation, Weka calls the training algorithm 11 times. Each data point is used once for testing and 9 times for training [31].

3.2. Classifier training and model storage
We have selected seven regression algorithms for testing. In assessing regression algorithms most significant are two parameters: correlation coefficient and root mean squared error. The correlation coefficient represents how well the predictions are correlated or change with the actual output value. A value of 0 is the worst and a value of 1 is a perfectly correlated set of predictions. Root mean squared error is the average amount of error made on the test set in the units of the output variable. This measure helps you get an idea of the amount a given prediction may be wrong on average [32]. For completeness
and possibility for the reproduction of the research in tables for each machine algorithm three more indicators are specified - mean absolute error, relative absolute error, and the mean value.

3.2.1. Using Zero Rule Regression MLA. Zero Rule Regression MLA is a basic method in the regression algorithm and predicts the mean of the training dataset [32, 33, 36]. The classifier is called ZeroR and is located in the rules group. The obtained results are in Table 2.

Table 2. Results when using Zero Rule Regression MLA.

|                         | full set | young after | old after | young before | old before |
|-------------------------|----------|-------------|-----------|--------------|------------|
| Correlation coefficient | -0.2578  | -0.4423     | -0.7121   | -0.3188      | -0.6923    |
| Mean absolute error     | 1.923    | 1.8276      | 1.891     | 1.6346       | 1.696      |
| Root mean squared error | 2.4941   | 2.2102      | 2.2825    | 1.987        | 2.0776     |
| Relative absolute error | 100 %    | 100 %       | 100 %     | 100 %        | 100 %      |
| Root relative squared error | 100 % | 100 %       | 100 %     | 100 %        | 100 %      |
| The mean value          | 25.2083  | 26.5711     | 26.5711   | 23.8455      | 23.8455    |

For the full set of training data, the ZeroR algorithm predicts the mean soil moisture value of 25.21 (in % volumetric water content) and root relative squared error of 2.49. Any other suitable and working MLA must achieve a value better than this.

3.2.2. Using Linear Regression MLA. In use since 1802. Works well if there are more instances than attributes [31]. Linear algorithms assume that the predicted attribute is in linear combination with the other input attributes. The classifier is called LinearRegression and is located in the function group. The obtained results are in Table 3.

Table 3. Results when using Linear Regression MLA.

|                         | full set | young after | old after | young before | old before |
|-------------------------|----------|-------------|-----------|--------------|------------|
| Correlation coefficient | 0.5729   | 0.4525      | 0.3777    | -0.3146      | -0.4479    |
| Mean absolute error     | 1.6567   | 1.6233      | 1.7705    | 1.6645       | 1.7672     |
| Root mean squared error | 2.0291   | 1.9665      | 2.0358    | 2.0259       | 2.1848     |
| Relative absolute error | 86.1492 %| 88.8204 %   | 93.628 %  | 101.8232 %   | 104.1985 % |
| Root relative squared error | 81.353 % | 88.9744 %   | 89.1881 % | 101.9594 %   | 105.1588 % |

As you can see there is a proven highly non-linear relationship between the studied parameters. Therefore, we move to nonlinear algorithms.

It should be noted here that there is a very high correlation coefficient between R and G color components - 0.9891 (R = 0.8772 * G -3.7652). Weka contains algorithms that eliminate the influence of highly correlated and completely uncorrelated variables.

3.2.3. Using Nonlinear k-Nearest Neighbors Regression MLA. The k-nearest neighbors algorithm [37, 38] proposed by Thomas Cover is a non-parametric method used for classification and regression. In both cases, the input consists of the k-closest training examples in the feature space. It looks for k-Nearest Neighbors and chooses the majority class among several neighbors [31]. The classifier is called IBk and is located in the lazy group. The obtained results are in Table 4.
The size of the neighborhood is controlled by the k parameter, called IBk in Weka. We use value for k = 7 because we get the best results from it.

3.2.4. Using Nonlinear Model Tree Regression MLA. The model is a tree, each leaf of which has a linear regression model [31]. The original algorithm M5 is invented by R. Quinlan and Yong Wang made improvements.

The classifier is called M5P and is located in the trees group. The obtained results are in Table 5.

### Table 4. Results when using k-Nearest Neighbors Regression MLA.

|                          | full set  | young after | old after | young before | old before |
|--------------------------|-----------|-------------|-----------|--------------|------------|
| Correlation coefficient  | 0.6948    | 0.7286      | 0.2674    | 0.3304       | -0.1916    |
| Mean absolute error      | 1.439     | 1.3447      | 1.781     | 1.5796       | 1.7414     |
| Root mean squared error  | 1.7846    | 1.5041      | 2.1217    | 1.9197       | 2.1082     |
| Relative absolute error  | 74.8305%  | 73.5788%    | 94.179%   | 96.631%      | 102.6749%  |
| Root relative squared error | 71.553%  | 68.0524%    | 92.9548%  | 96.6124%     | 101.4741%  |

Here you can see the view of the tree in folded form:

Soil_Temp <= 20.08 :
| Soil_Temp <= 19.2 :
| | | Soil_Temp <= 19.08 :
| | | | | Soil_Temp <= 18.605 : LM1 (9/0%)
| | | | | Soil_Temp >  18.605 : LM2 (27/68.949%)
| | | | Soil_Temp >  19.08 : LM3 (18/0%)
| | | Soil_Temp >  19.2 : LM4 (18/0%)
| | Soil_Temp >  20.08 :
| | | Soil_Temp <= 21.32 : LM5 (63/47.358%)
| | | | | Soil_Temp >  21.32 :
| | | | | | Soil_Temp <= 21.51 : LM6 (9/0%)
| | | | | | | Soil_Temp >  21.51 : LM7 (18/0%)

The nonlinear model is divided into seven linear ones:
LM num: 1
Soil_Mois = 0.0185 * G - 2.9332 * Soil_Temp + 76.6341

LM num: 2
Soil_Mois = 0.0135 * G - 6.1927 * Soil_Temp + 138.9254

LM num: 3
Soil_Mois = 0.0087 * G + 1.8729 * Soil_Temp - 13.2783

LM num: 4
Soil_Mois = 0.0064 * G + 0.7358 * Soil_Temp + 7.475

LM num: 5
Soil_Mois = -0.0463 * R + 0.0508 * B + 0.759 * Soil_Temp + 11.0926

LM num: 6
Soil_Mois = -0.0176 * R + 0.018 * B + 6.9154 * Soil_Temp - 121.964

LM num: 7
Soil_Mois = -0.0176 * R + 0.018 * B + 11.0999 * Soil_Temp - 211.6938

Figure 2 is a visual representation of the tree:

![Tree Diagram](image)

**Figure 2.** Visualization of resulting tree after testing with M5P algorithm.

3.2.5. Using Nonlinear Decision Tree Regression MLA. Decision trees work by creating a tree to evaluate an instance of data.

Decision tree learning is predictive modeling uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).

The classifier is called REPTree and is located in the trees group. The obtained results are in Table 6.
Table 6. Results when using Decision Tree Regression MLA.

|                     | full set | young after | old after | young before | old before |
|---------------------|----------|-------------|-----------|--------------|------------|
| Correlation coefficient | 0.8915 | 0.817       | 0.6034   | 0.6477       | 0.2294     |
| Mean absolute error   | 0.7237  | 0.9252      | 1.5235   | 1.0161       | 1.6712     |
| Root mean squared error | 1.1232 | 1.2534      | 1.7343   | 1.5144       | 1.9615     |
| Relative absolute error | 37.6321 % | 50.6212 % | 80.5669 % | 62.1589 % | 98.5389 % |
| Root relative squared error | 45.0343 % | 56.7128 % | 75.9796 % | 76.2157 % | 94.4101 % |
| Size of the tree (rules) | 35 | 5 | 5 | 19 | 7 |

The tree is inverted, starting at the top or root of the tree and moving down to the leaves until a prediction can be made [32].

3.2.6. Using Nonlinear Support Vector Regression MLA. SVR works by finding a line of best fit that minimizes the error of a cost function. These instances are called support.

The method of support vectors (Support Vector Machines - SVM) was proposed by Vladimir Vapnik in 1963 as a method for constructing an optimal linear classifier [44]. The requirement for linear class separation limits the method’s application. The real development of the method took place after 1992, when was proposed a generalized version [45], enabling the method to be used for nonlinear problems as well. SVM is a method of classification through training with a teacher. The classifier is called SMOreg and is located in the function group. The obtained results are in Table 7.

Table 7. Results when using Support Vector Regression MLA.

|                     | full set | young after | old after | young before | old before |
|---------------------|----------|-------------|-----------|--------------|------------|
| Correlation coefficient | 0.574 | 0.4997       | 0.3567   | -0.2667     | -0.6186    |
| Mean absolute error   | 1.6071  | 1.5343      | 1.7544   | 1.6089       | 1.7353     |
| Root mean squared error | 2.0595 | 1.9016      | 2.0475   | 2.1274       | 2.2339     |
| Relative absolute error | 83.5732 % | 83.9533 % | 92.7738 % | 98.422 % | 102.3181 % |
| Root relative squared error | 82.5741 % | 86.0408 % | 89.7042 % | 107.0673 % | 107.5216 % |

3.2.7. Using Nonlinear Multilayer Perceptron Regression MLA [46]. The classifier is called MultilayerPerceptron and is located in the function group. The obtained results are in Table 8.

It is also called artificial neural networks or simply neural networks for short. Neural networks are a complex algorithm to use for predictive modeling because there are so many configuration parameters that can only be tuned effectively through intuition and a lot of trial and error.

Table 8. Results when using Multilayer Perceptron Regression MLA.

|                     | full set | young after | old after | young before | old before |
|---------------------|----------|-------------|-----------|--------------|------------|
| Correlation coefficient | 0.629 | 0.6468       | 0.4663   | 0.4117       | 0.3309     |
| Mean absolute error   | 1.6006  | 1.3921      | 1.5978   | 1.5574       | 1.8266     |
| Root mean squared error | 1.9613 | 1.7627      | 2.3078   | 1.9168       | 2.1248     |
| Relative absolute error | 83.2332 % | 76.1691 % | 84.4912 % | 95.2763 % | 107.7025 % |
| Root relative squared error | 78.6368 % | 79.7565 % | 101.1056 % | 96.4683 % | 102.2746 % |
3.2.8. Summarize table. In table 9 is presented in a short sum of the above calculations.

**Table 9. Summary table.**

|                     | full set | ZeroR | Linear Regression | KNN - 7 | M5P | REPTree | SMOreg | MultyPerc |
|---------------------|----------|-------|-------------------|---------|-----|---------|--------|-----------|
| Correlation coefficient | -0.2578 | 0.5729 | 0.6948 | 0.8587 | 0.8915 | 0.574 | 0.629 |
| Mean absolute error | 1.923 | 1.6567 | 1.439 | 1.0556 | 0.7237 | 1.6071 | 1.6006 |
| Root mean squared error | 2.4941 | 2.0291 | 1.7846 | 1.3151 | 1.1232 | 2.0595 | 1.9613 |
| Relative absolute error | 100 % | 86.1492 % | 74.8305 % | 54.8902 % | 37.6321 % | 83.5732 % | 83.2332 % |
| Root relative squared error | 100 % | 81.353 % | 71.553 % | 52.726 % | 45.0343 % | 82.5741 % | 78.6368 % |
| Number of Rules (Leaves) | 7 | 35 |

3.2.9. Testing of the selected and stored model with the additional data set provided for this purpose, not included in the stored model

Although the Decision Tree has a high correlation coefficient (0.8915) we decided to check the operability of the stored model with D5P MLA (0.8587) because in Decision Tree you get a lot of leaves or rules – 35, and with D5P there are only 7. Table 10 presents the resulting prediction error and the corresponding percentage, using the stored M5P model. The results are more than positive:

**Table 10. Sample from the obtained data during approbation of the selected model.**

| Inst | actual | predicted | error | error, % | Leaf  | Inst | actual | predicted | error | error, % | Leaf  |
|------|--------|-----------|-------|----------|-------|------|--------|-----------|-------|----------|-------|
| 1    | 22     | 22.908    | 0.908 | 3.96%    | young | 14   | 23.21  | 23.705    | 0.495 | 2.09%    | young |
| 2    | 22     | 23.137    | 1.137 | 4.91%    | young | 15   | 23.21  | 23.975    | 0.765 | 3.19%    | young |
| 3    | 22     | 22.72     | 0.72  | 3.17%    | young | 16   | 23.21  | 23.732    | 0.522 | 2.20%    | young |
| 4    | 22     | 22.72     | 0.72  | 3.17%    | young | 17   | 23.21  | 23.867    | 0.657 | 2.75%    | young |
| 5    | 22     | 22.895    | 0.895 | 3.91%    | young | 18   | 21.45  | 23.641    | 2.191 | 9.27%    | old   |
| 6    | 22     | 22.72     | 0.72  | 3.17%    | young | 19   | 21.45  | 23.372    | 1.922 | 8.22%    | old   |
| 7    | 22     | 22.895    | 0.895 | 3.91%    | young | 20   | 21.45  | 23.802    | 2.352 | 9.88%    | old   |
| 8    | 22     | 23.258    | 1.258 | 5.41%    | young | 21   | 21.45  | 23.722    | 2.272 | 9.58%    | old   |
| 9    | 23.21  | 23.261    | 0.051 | 0.22%    | young | 22   | 21.45  | 23.789    | 2.339 | 9.83%    | old   |
| 10   | 23.21  | 23.45     | 0.24  | 1.02%    | young | 23   | 21.45  | 24.125    | 2.675 | 11.09%   | old   |
| 11   | 23.21  | 23.611    | 0.401 | 1.70%    | young | 24   | 21.45  | 24.112    | 2.662 | 11.04%   | old   |
| 12   | 23.21  | 23.369    | 0.159 | 0.68%    | young | 25   | 21.45  | 24.327    | 2.877 | 11.83%   | old   |
| 13   | 23.21  | 23.49     | 0.28  | 1.19%    | young | 26   | 21.45  | 24.476    | 3.026 | 12.36%   | old   |

3.3. Comparison of another statistical method and machine learning methods

In our previous study [10] we also used soil temperature and leaf color to predict soil moisture. There we got the best results (lowest prediction error) also using nonlinear estimation (Piecewise linear regression with breakpoint [40-43]). In the same study, we unsuccessfully tried to model the above dependence with linear estimations (multiple linear regression and second-degree polynomial regression).

A comparison between this preliminary statistical survey [10] and the current one is presented in the form of Table 11.
Table 11. Another statistical and machine learning methods comparison

|                          | Another Statistical method [10]                                      | Machine learning method |
|--------------------------|-----------------------------------------------------------------------|-------------------------|
| Estimation               | Nonlinear                                                             | Nonlinear               |
| Algorithm                | Piecewise linear regression with breakpoint                         | MP5                     |
| Correlation coefficient  | 93 – 97 %                                                             | 85 %                    |
| The nonlinear model is   | Into two linear equations depending on the breakpoint                | Into seven linear equations depending on soil temperature |
| separated                |                                                                       |                         |
| Prediction error by young leaves | -4.85% to +14.98%                                                     | 0-5%                    |
| Prediction error by old leaves | too high                                                             | 8-12%                  |

The advantages of the Piecewise linear regression with breakpoint are: 1) the nonlinear model is reduced to only two linear equations and 2) the high correlation coefficient between input parameters and the predicted one. The disadvantage is that it does not work in predicting soil moisture based on old leaves.

In the MP5, the nonlinear dependence is described by seven linear equations and works on both young and old leaves with a low error rate.

4. Conclusions

The following seven classifiers for regression MLA with five training data sets were trained:
- Zero Rule;
- Linear Regression;
- k-Nearest Neighbors;
- M5P Nonlinear Model Tree;
- Decision Tree;
- Support Vector;
- Multilayer Perceptron.

The comparison of the obtained results by MLA indicates the following: The best result was obtained with D5P. Although Decision Tree has a high correlation coefficient (0.8915 we chose to check the performance with D5P (0.8587), because Decision Tree has a large number of rules (35), and in D5P there are only 7 (Table 9). Table 10 presents the resulting prediction error and the corresponding percentage, using the stored M5P model.

The comparison of D5P MLA with Piecewise linear regression with breakpoint (Table 11) gives us reason to conclude, that both methods have their advantages and disadvantages, but both are very effective in the resulting nonlinear relationship – in Piecewise linear regression with breakpoint, the nonlinear dependence was reduced to only two linear equations, depending on the breakpoint, and in D5P they are seven. D5P predicts very satisfactory soil moisture in both young and old leaves, while Piecewise linear regression with breakpoint applies only to the young leaves before watering. In D5P is obtained an error rate of 0-5% in predicting soil moisture based on the color of young leaves and 8-12% based on the color of the old ones before watering, taking into account the soil temperature in the model.

Finally, this proves again the preliminary hypothesis that the young leaves before watering are the best indicator of the need for irrigation.

References
[1] Koumanov K, Matev A, Kornov G and Germanova D 2018 Irrigation management in perennial crops by sensing the soil water status Journal of Mountain Agriculture on the Balkans 21(5) pp 182-192
[2] Koumanov K, Matev A, Kornov G and Germanova D 2018 Irrigation management in perennial crops by sensing the plant water status Journal of Mountain Agriculture on the Balkans 21(5) pp 172-181
[3] Dobriyal P, Qureshi A, Badola R and Hussain S A 2012 A review of the methods available for estimating soil moisture and its implications for water resource management Journal of Hydrology 458 pp 110-117

[4] Atanasov S 2015 Soil specific FDR sensor calibration in soil moisture measuring Research Papers of Rousse University 54 pp 217-221

[5] Bianchi A, Masseroni D, Thalheimer M, de Medici L O and Facchi A 2017 Field irrigation management through soil potential measurements: a review Italian Journal of Agrometeorology 22(2) pp 25-38

[6] Pardossi A, Incrocci L, Incrocci G, Malorgio F, Battista P, Bacci L and Balendonck J 2009 Root zone sensors for irrigation management in intensive agriculture Sensors 9(4) pp 2809-2835

[7] Turner N 1988 Measurement of plant water status by the pressure chamber technique Irrig Sci 9 pp 289–308

[8] Kalaidjieva R, Kuneva V and Matev A 2015 Using on infrared thermometer for establishment of water stress in green beans irrigated by different pre-irrigation soil moisture Plant Science 5 pp 72-78

[9] Sela E, Cohen Y, Alchanatis V, Saranga Y, Cohen S, Möller M and Orolov V 2007 Thermal imaging for estimating and mapping crop water stress in cotton European conference in precision agriculture Wageningen Academic Publishers pp 365-371

[10] Atanasov S S, Daskalov P I and Nedeva V I 2016 An intelligent approach of determining relationship between tomato leaves color and soil moisture and temperature Bulgarian Journal of Agricultural Science 22(6) pp 1027-1035

[11] Guo D, Juan J, Chang L, Zhang J and Danfeng Huang 2017 Discrimination of plant root zone water status in greenhouse production based on phenotyping and machine learning techniques Scientific Reports 7(1) pp 1-12

[12] Mekonnen Y, Namuduri S, Burton L, Sarwat A and Bhansali S 2019 Machine learning techniques in wireless sensor network based precision agriculture Journal of the Electrochemical Society 167(3)

[13] Ma C, Zhang H H and Wang X 2014 Machine learning for big data analytics in plants Trends in plant science 19(12) pp 798-808

[14] Virmodkar S S, Pachghare V K, Patil V C and Jha S K 2020 Remote sensing and machine learning for crop water stress determination in various crops: a critical review Precision Agriculture pp 1-35

[15] Alhnaity B, Pearson S, Leontidis G and Kollias S 2019 Using deep learning to predict plant growth and yield in greenhouse environments arXiv, arXiv:1907

[16] Mekonnen Y, Namuduri S, Burton L, Sarwat A and Bhansali S 2019 Machine learning techniques in wireless sensor network based precision agriculture Journal of the Electrochemical Society 167(3)

[17] Gutiérrez S, Diago M P, Fernández-Novales J and Tardaguila J 2018 Vineyard water status assessment using on-the-go thermal imaging and machine learning PloS One 13(2)

[18] Patil H 2019 Regression test suite study using classic statistical methods and machine learning

[19] Afzal A, Duiker S W, Watson J E and Lucthe D 2017 Leaf thickness and electrical capacitance as measures of plant water status Transactions of the ASABE 60(4) pp 1063-1074

[20] Afzal A, Duiker S W and Watson J E 2017 Leaf thickness to predict plant water status Biosystems Engineering 15 pp 148-156

[21] Mahan J R, Young A W and Payton P 2015 Continuously monitored canopy temperature as a proxy for plant water status American Journal of Plant Sciences 6(14) p 2287

[22] Ko-Madden C T, Upadhyaya S K, Kizer E E, Drechsler K M, Rojo F, Meyers J N and Schramm A E 2017 Precision irrigation in wine grape using a proximal leaf monitor system for measuring plant water status ASABE Annual International Meeting (p. 1) American Society of Agricultural and Biological Engineers

[23] Atanasov S 2017 Monitoring the time of irrigation water absorption by in indeterminate tomato
plants based on colorimetric measurements of the leaves University of Ruse, Proceedings of the Student Scientific Session - SSS’17 pp 64-69 (Bg)

[24] Dhillon R S 2015 Development and evaluation of continuous leaf monitoring system for measurement of plant water status (Doctoral dissertation) University of California, Davis

[25] Rojo F, Kizer E, Upadhyaya S, Ozmen S, Ko-Madden C and Zhang Q 2016 A leaf monitoring system for continuous measurement of plant water status to assist in precision irrigation in grape and almond crops IFAC-PapersOnLine 49(16) pp 209-215

[26] Virk A L, Noor M A, Fiaz S, Hussain S, Hussain H A, Rehman M, Ahsan M and Ma W 2020 Smart farming: an overview Smart Village Technology Springer, Cham pp 191-201

[27] Nishina H 2015 Development of speaking plant approach technique for intelligent greenhouse Agriculture and Agricultural Science Procedia 3(9) p 13

[28] Atanasova P 2020 Effect of innovations on the economic development of the Yambol region Trakia University pp 68-82

[29] Ovcharova A and Harizanova-Petrova B 2012 Influence of the water factor on the growth and development of celery in the region of Plovdiv Research Papers of Rousse University 51 pp 121-125 (Bg)

[30] Popova Z and Ivanova M 2012 Impact of climate uncertainties and soil characteristics on maize yield and irrigation requirements in Plovdiv region Water Managements 5/6 pp 26-33 (Bg)

[31] Witten I H 2013 Data mining with weka Department of Computer Science University of Waikato New Zealand

[32] Brownlee J 2016 Machine learning mastery with Weka ebook edition v. 1.1

[33] Eibe F, Hall M and Witten I H 2016 The WEKA workbench. Online appendix for "Data mining: practical machine learning tools and techniques" Morgan Kaufmann

[34] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P and Witten I H 2009 The WEKA Data Mining Software: An Update SIGKDD explorations 11

[35] University of Waikato “Weka Manual for version 3-9-3” available online at University website

[36] Bouckaert R, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A and Scuse D 2019 WEKA manual for version 3-8-4 University of Waikato Hamilton New Zealand

[37] Cover T and Hart P 1967 Nearest neighbor pattern classification IEEE Transactions on Information Theory 13(1) 21-27

[38] Altman N S 1992 An introduction to kernel and nearest-neighbor nonparametric regression The American Statistician 46(3) 175–185

[39] Mendoza M R N and Ramirez M M 2019 Automatización de riego y fertilización mediante el uso de red inalámbrica de sensors I Foro Internacional sobre Sistemas de Producción Agropecuarios Mecanizados y Computarizados, Calzeta, Ecuador (Esp)

[40] Nisbet R, Elder J and Miner G 2009 Handbook of statistical analysis and data mining applications Academic Press

[41] James G, Hastie T, Tibshirani R, Witten D 2014 An Introduction to Statistical Learning: With Applications in R Springer

[42] Yang L, Liu S, Tsopa S and Papageorgiou L G 2016 Mathematical programming for piecewise linear regression analysis Expert systems with applications 44 pp. 156–167

[43] Malash G F and El-Khiairy M I 2010 Piecewise linear regression: A statistical method for the analysis of experimental adsorption data by the intraparticle-diffusion models Chemical Engineering Journal 163(3) pp 256-263

[44] Vapnik V 1995 The Nature of Statistical Learning Theory John Wiley & Sons, New York

[45] Boser B E, Guyon I M and Vapnik V N 1992 A training algorithm for optimal margin classifiers Proceedings of the fifth annual workshop on Computational learning theory pp. 144-152

[46] Haykin, S 2008 Neural Networks and Learning Machines 3rd ed Prentice Hall Upper Saddle River