KSÜ Tarım ve Doğa Derg 27 (2), 443-452, 2024
KSU J. Agric Nat 27 (2), 443-452, 2024
DOI:10.18016/ksutarimdoga.vi.1215856

Comparison of Predictive Performance of Data Mining Algorithms in Predicting Tomato Yield with the A Case Study in Igdir

Köksal KARADAŞ[1✉], Osman Doğan BULUT[2]
[1,2] Department of Agricultural Economics, Faculty of Agriculture, Igdir University, Iğdır, Türkiye
[1]https://orcid.org/0000-0003-1176-3313, [2]https://orcid.org/0000-0003-2682-6356
✉: koksal.karadas@igdir.edu.tr

ABSTRACT
Among the vegetable species in the world, the plant with the most cultivation area is tomato. Increasing tomato yield is important in terms of contributing more to the world economy and farmer's income. With the advancement in software technologies, the importance of data mining algorithms is increasing due to the fact that these algorithms can produce more sophisticated solutions for regression and classification problems. Determining the factors affecting tomato yield and comparing different data mining algorithms on prediction of tomato yield are the purpose of this study. For this purpose, survey study was conducted with the 105 farmers in Igdir province. Different data mining algorithms including Classification and Regression Tree (CART), Chi-Square Automatic Interaction Detector (CHAID), Exhaustive CHAID, Artificial Neural Network Algorithm (ANN), Multivariate Adaptive Regression Splines (MARS) and General Linear Model (GLM) were developed and compared their predictive performance. MARS decision tree built a model with greatest predictive accuracy. The superiority order in the predictive accuracy of the other algorithms were ANN> GLM> CART> CHAID> Exhaustive CHAID. In the MARS model, number of irrigation, amount of chemical fertilizer, age of farmer, number of seedlings, education level, soil analysis status, sowing region were found statistically significant (P<0.05). Preferring the MARS model could allow detecting factors affecting tomato yield and their interactions with higher accuracy. To increase yield, at least 1450 seedlings should be planted per decare and irrigation should be at least 5 times.

Agricultural Economics

Research Article

Article History
Received : 07.12.2022
Accepted : 07.09.2023

Keywords
Data mining algorithms
Production economics
Tomato yield
Igdir

Farklı Veri Madenciliği Algoritmalarının Domates Verimindeki Tahmin Performanslarının Karşılaştırılması: Iğdır İli Örneği

ÖZET
Domates sebze türleri arasında en fazla ekim alanına sahip bitkidir. Domates veriminin artırılması dünya ekonomisi ve çiftçi gelirine daha fazla katkı sağlaması açısından önemlidir. Yazılım teknolojilerinin ilerlemesi ile regresyon ve sınıflandırma problemlerine daha gelişmiş çözümlerin sunulması veri madenciliğinin önemi artırmaktadır. Bu çalışmada domates verimini etkileyen faktörlerin belirlenmesi ve domates veriminin tahmininde farklı veri madenciliği algoritmalarının karşılaştırılması amaçlanmıştır. Bu amaç ile Iğdır ilinde 105 çiftçi ile anket çalışması yapılmıştır. Sınıflandırma ve Regresyon Ağacı (CART), Ki-Kare Otomatik Etkileşim Dedektörü (CHAID), Exhaustive CHAID, Yapay Sinir Ağı Algoritması (ANN), Çok Değişkenli Uyarlamalı Regresyon Analizi (MARS) ve Genel Doğrusal Model (GLM) gibi farklı veri madenciliği algoritmaları kullanılarak tahmin performansları karşılaştırılmıştır. MARS karar ağacı, en yüksek tahmin doğruluğuna sahip modeli oluşturmuştur. Tahmin performanslarına göre diğer algoritmalar ANN> GLM> CART> CHAID> Exhaustive CHAID'dır. MARS modelinde, sulama sayısı, kimyasal gübre miktarı, çiftçi yaşı, fide sayısı, eğitim düzeyi,

Tarım Ekonomisi

Araştırma Makalesi

Makale Tarihçesi
Geliş Tarihi : 07.12.2022
Kabul Tarihi : 07.09.2023

Anahtar Kelimeler
Veri madenciliği algoritmaları
Üretim ekonomisi
Domates verimi
Iğdır

toprak analiz durumu ve ekim bölgesi değişkenleri istatistiksel olarak anlamlı bulunmuştur (P<0.05). MARS modelinin tercih edilmesi, domates verimini etkileyen faktörleri ve bunların etkileşimlerini daha yüksek doğrulukla tespit edilmesini sağlayacaktır. Verim artışı için dekara en az 1450 fide dikilmeli ve en az 5 defa sulama yapılmalıdır.

Atıf İçin: Karadaş, K., & Bulut, O.D., (2024). Farklı Veri Madenciliği Algoritmalarının Domates Verimindeki Tahmin Performanslarının Karşılaştırılması: Iğdır İli Örneği. *KSÜ Tarım ve Doğa Derg 27*(2), 443-452. https://doi.org/10.18016/ksutarimdoga.vi.1215856

To Cite: Karadaş, K., & Bulut, O.D., (2024). Comparison of Predictive Performance of Data Mining Algorithms in Predicting Tomato Yield with the A Case Study in Igdir. *KSU J. Agric Nat 27*(2), 443-452. https://doi.org/10.18016/ksutarimdoga.vi.1215856

INRODUCTION

Tomato, which has the largest cultivation area among vegetable species in the world, contains vitamins A, B1, B2, C, and K and essential amino acids, sugars and dietary fibers and it is easy to digest and is very rich in minerals (Sajjad et al., 2011; Debela et al., 2016; Özkan et al., 2017; Tatar & Pirinç, 2017). Among the main benefits of tomatos that can grow in a wide climate zone, it is well known that tomato has a positive effect on diagnosing some chronic and cardiovascular diseases and prevents cancer, prostate and liver fat (Tapiero et al., 2004; Navarro-González et al., 2018). It is also an important antioxidant thanks to its lycopene content (Söylemez & Pakyürek, 2017). Tomato, which can be consumed fresh, is also used as raw material in ketchup and canned food production, fruit juice industry, dried and frozen consumption and in the fruit and vegetable industry (Manan et al., 2016). The best daytime temperature for tomato growth is 21-24 ° C, and the ideal temperature for fruit set and pollination is 24 ° C and 17 ° C day and night (Comlekcioglu & Simsek, 2014).

Around the World, 182,256,458 tons of tomatoes were produced in an area of 40,762,457 da and tomato yield is 3,827 kg da-1. Turkey ranks fourth in the total amount of production (Anonymous, 2018). Although the tomato yield is 7,414 kg da-1 in Turkey, this value decreases to 3,470 kg da-1 in Igdir province. Tomato production amount in Igdir province is 33,732 tons year-1.

The tomato yield level in Igdir province, which has the appropriate climate and soil conditions for tomato production, is lower than half the average of Turkey (Anonymous, 2019; Anonymous, 2020). To increase the tomato yield level, it is important to determine the factors affecting the tomato yield and to develop solutions that will increase the yield.

Some of the studies on tomato production and yield; Hahn (2013) stated that controlled fertilization and irrigation increased the income of the producer by saving water and fertilizer as well as optimizing the yield of tomatoes, Özer (2016) found out that the use of quality seeds and seedlings increases the yield, Kibria et al. (2016) pointed out that biogas production residues are an alternative to chemical fertilization in tomato yield increase, Tesafay et al. (2018) disclosed that 50% vermicompost and 50% mineral fertilizer provide more economical production with increased yield in tomato production, Liu et al. (2019) proposed that excessive K fertilization of tomatoes during the fruit maturity period with adequate irrigation reduces the yield. Regarding high value-added agricultural business that is high-quality and high-yielding cultivation techniques, some factors which are humidity, water per m^2, receiving light amount, phosphite were also put forward by some scholars (Estrada-Ortiz et al., 2012; Letourneau et al., 2015; Na et al., 2017; Cho et al., 2018).

In recent years, some prediction models have been used to predict, evaluate and classify the agricultural activity results. Data mining is one of these techniques that is widely used for classification and estimation in many fields such as engineering, marketing strategy and industry, and its use in the agricultural field is very limited (Camdeviren et al., 2007). While there are more studies on data mining, especially in the field of animal husbandry (Aytekin et al., 2018; Celik et al., 2018; Karadas & Birinci, 2019), these studies are quite limited in crop production (Küçükönder et al., 2015; Irmak & Ercan, 2017; Bostanci & Eren-Atay, 2018). The aim of this study is to determine the various factors affecting tomato yield by employing data mining algorithms. In addition, determining the algorithm with the highest predictive power among; Classification and Regression Trees (CART), Exhaustive CHAID, Chi-Square Automatic Interaction Detector (CHAID), Multivariate Adaptive Regression Splines (MARS), General Linear Model (GLM) and Artificial Neural Network Algorithm (ANN) have been focus of this paper.

MATERIALS and METHODS

Materials

Igdir province, is located on the easternmost border of Turkey with 3 neighboring countries of Armenia, Nakhchivan and Iran (Figure 1). Igdir is located between 390 55' latitude and 440 03' longitude and is known as 850 m above sea level. The data obtained from the survey conducted with face-to-face interviews with 105 farmers producing tomatoes in Igdir province

is the main material of this study. The survey study was conducted between August to September 2016 after the tomato harvest and the study covers the 2016 production period. In addition to the survey; previous studies, reports and statistical data of various organizations on the subject were also utilized to support and validate the outcome of this work.

Figure 1. The location of Igdir province in Turkey
Şekil 1. Iğdır ilinin Türkiye'deki konumu

The central district and Karakoyunlu district, which has more than 95% of the total tomato production in Igdir, were selected as the research area. The information on the agricultural businesses producing tomatoes in these regions has been obtained from the Agricultural Institutions of Turkey. Using the Simple Random Sampling method (Yamane, 2010), the sample size was calculated as 95 agricultural businesses engaged in tomato production (90% confidence level and 10% deviation). Due to the possibility of having incorrect or missing data, the number of questionnaires was increased by 10% and the sample volume was increased to 105. The sampling equation is given below.

$$n = \frac{NS^2}{(N-1)D^2 + S^2} \qquad (1.)$$

n: The number of agricultural business engaged in tomato production to represent the population

N: The total number of agricultural business engaged in tomato production (465)

S^2: Population variance (33.17)

D: Refers to the impact factor

The correction factor (D) = $(E/t)^2$ was obtained from the equation, and the t coefficient was taken as 1.6445 for 90% confidence. E is the error (0.87), it is 10% of the average size group.

The distribution of the survey quantities by regions was shown in Table 1.

Table 1. Survey quantities by districts
Çizelge 1. Anket sayılarının bölgelere göre dağılımı

District	Agricultural Business	Sample Size	Percentage (%)
Center	239	55	52.6
Karakoyunlu	217	50	47.4
Total	456	105	100

Methods

In the created models, the dependent variable is TY (tomato yield-as kg ha^{-1}). Independent variables are NI (number of irrigation), ACF (amount of chemical fertilizer-kg ha^{-1}), AM (amount of medicine-ml), AF (age of farmer), NS (number of seedlings), SD (sowing date), EL (education level: illiterate = 1, literate = 2, primary school = 3, secondary school = 4, high school = 5, associate degree = 6, undergraduate=7), SAS (soil analysis status 1 = yes, 2 = no), SR (sowing region-1 = center, 2 = Karakoyunlu) and AFM (amount of farm manure kg ha-1).

Although one-way analysis of variance (ANOVA) is used in many fields, it may cause misleading results in cases where some assumptions are violated. The CART, Exhaustive CHAID and CHAID algorithms are effectively used to create models in nominal, ordinal and scale variables, and the CART algorithm allows to create a decision tree structure based on binary split criteria by dividing a node repeatedly into two sub-nodes (Duru et al., 2017; Eyduran et al., 2017).

The more successful the division is, the more similarities arise between the members of the outcome groups (Sun & Hui, 2008). The number of producers in the main and sub nodes was determined as 8:4 to obtain the highest prediction performance of TY algorithms. In the SPSS program, the pruning option is enabled to remove unnecessary nodes in the CART algorithm, unlike the CHAID and Comprehensive CHAID algorithms, which create multiple split nodes so that the variance within the nodes is minimal (Karadas & Kadirhanogullari, 2017). Since TY is a continuous variable, the F test was used to check the significance of the effective independent variables in CHAID algorithms. The General Linear Model (GLM) using the Least Squares Method and Artificial Neural Networks (ANN), which is multi-layered and resembling the human brain, have been used in many studies to determine the predictive power of the model (Duru et al., 2017; Karadas & Kadirhanogullari, 2017; Eyduran et al., 2017;). The MARS data mining algorithm, a nonparametric regression method, allows the use of piecewise basic functions to define a response variable and a set of input variables. It automatically determines the node locations and can be shown in the following equation (Eyduran et al., 2017).

$$f_M(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x) \qquad (2.)$$

Basic function parameters of the MARS algorithm are ß0 and ßm. Bm (x), which is the spline basis function, is used as follows:

$$B_m(x) = \prod_{k=1}^{k_m} \left[s_{km} \left(x_{v(k,m)} - t_{k,m} \right) \right] \qquad (3.)$$

In the equation, km is takes the number of nodes and Skm takes values either -1 or 1 and indicates the right and left boundaries of the function. v(k,m) indicates the label of the input variable and tk,m indicates the location of the node (Friedman 1991).

Generalized cross validation (GCV) eliminates unnecessary basic functions.

$$GCV = \frac{\frac{1}{N}\sum_{i=1}^{N}\left[y_1 - \hat{f}(x_i)\right]^2}{\left[1 - \frac{c(B)}{N}\right]^2} \qquad (4.)$$

N: the number of data

c (B): a complexity penalty

CART, CHAID, MARS and all other algorithms contain significant variables and provide information about estimators in studies. In the statistical analysis of the data, TY defined as dependent variable. Strong predictions of CHAID, CART, Exhaustive CHAID, MARS algorithms and MLP, which is the application of ANN, were compared. CART, CHAID, MARS and all other algorithms contain important variables and provide information about estimators in studies. What is important for the scientist is to determine the effect of independent variable, which are the predictors, on the dependent variable and to reveal the degree of their interaction. Model selection criteria compared by performance are given:

Coefficient of Determination (R^2):

$$R^2(\%) = \left[1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} \right] * 100 \qquad (5.)$$

Adjusted Coefficient of Determination (Adj. R^2):

$$Adj.R^2(\%) = \left[1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} \right] * 100 \qquad (6.)$$

Coefficient of Variation (CV):

$$CV(\%) = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\varepsilon_i - \bar{\varepsilon}\right)^2}}{\bar{Y}} * 100 \qquad (7.)$$

Standard Deviation Ratio (SD):

$$SD_{ratio} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(\varepsilon_i - \bar{\varepsilon}\right)^2}{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}} \qquad (8.)$$

Relative Approximation Error (RAE):

$$RAE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n} Y_i^2}} \qquad (9.)$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n}} \qquad (10.)$$

Mean Absolute Deviation (MAD):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_{ip}\right| \qquad (11.)$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_{ip}}{y_i}\right| \cdot 100 \qquad (12.)$$

In the equations;

n: the number of sample in the population

k: the number of model parameters

yi (TY): observed actual values of the output variable

yip: TY estimation values

ε: Error term

IBM SPSS 23 package program was used for statistical evaluations of CART, CHAID, Exhaustive CHAID, ANN and GLM. STATISTICA 8.0 trial version was used in MARS algorithm.

RESULTS and DISCUSSION

A comparison of three data mining algorithms through some independent variables in the estimation of TY was first documented. For these data mining algorithms, the model evaluation criteria and results are given in Table 2. Superiority order in the predictive accuracy has been determined according to the R, R^2, Adj.R^2, value of which are requested to be high, and also; RMSE, RAE, CV (%), SD_{RATIO}, MAD, MAPE, the value of which are requested to be low. The order of superiority in the prediction accuracy of the algorithms was found as MARS > ANN > GLM > CART > CHAID > Ex.CHAID. The prediction performance of the MARS algorithm was found to be more advantageous than other algorithms in terms of selection criteria.

Table 2. Model evaluation criteria and results for data mining algorithms
Çizelge 2. Model değerlendirme kriterleri ve veri madenciliği algoritma sonuçları

Algorithm	r	R^2	Adj. R^2	RMSE	RAE	CV(%)	SD_{RATIO}	MAD	MAPE
EX. CHAID	0.682[b]	0.465	0.449	1112	0.1920	20.00	0.7311	833.671	0.1588
CHAID	0.696[b]	0.484	0.458	1092	0.1887	19.65	0.7184	790.514	0.1528
CART	0.710[b]	0.504	0.479	1071	0.1850	19.27	0.7043	815.280	0.1539
GLM	0.742[b]	0.551	0.528	1019	0.1760	18.33	0.6702	803.348	0.1489
ANN	0.751[b]	0.564	0.537	1005	0.1737	18.09	0.6613	783.885	0.1473
MARS	0.848[a]	0.719	0.679	806	0.1393	14.51	0.5301	614.612	0.1184

Factors affecting tomato yield in the comprehensive CART algorithm used to determine the results of the decision tree structure was given in Figure 2. Among the factors examined in the regression decision tree structure for CART algorithm, the number of seedlings (NS), number of irrigation (NI), education level (EL), fertilizer amount (ACF) and pesticide amount (AM) were determined as significant variables. In the CART algorithm, the determination coefficient was estimated as 71%.

While the highest tomato yield was obtained from Node 6 (8011 kg/da), the lowest tomato yield was obtained from Node 7 (4092 kg/da). The average tomato yield was 5,585 kg da^{-1} at the Node 0. This amount is higher than previous tomato yield studies; The world average yield level was 3,287 kg da^{-1} (Anonymous, 2018); moreover, 2,900 kg da^{-1} in the study of Neta et al. (2019) from Brazil and 3,158 kg da^{-1} in the study of Degefa et al. (2019) from Ethiopia. According to researches carried out in Turkey, tomato yield in Cukurova region is 5,812 kg da^{-1}, 7,602 kg da^{-1} in Isparta province and Turkey's average is 7,414 kg da^{-1} (Yaraş & Daşgan, 2012; Kiracı & Karataş, 2015; Anonymous, 2019). Various academic studies on tomato yield are available in the literature. However, finding effective independent variables using data mining algorithms to model tomato yield has not been done before. In this respect, this study will be the first in data mining applications.

The yield order between Nodes 1-2 was found as Node 1 (NS ≤ 1342) < Node 2 (NS > 1342). While a yield of 5,135 kg da^{-1} was obtained in Node 1, a yield of 6,659 kg da-1 was obtained in Node 2. It is understood that the producers need to plant more than 1342 seedlings per decare. Node 1 is divided into Node 3 (NI≤7.5) and Node 4 (NI> 7.5) sub-nodes. Node 4 was first terminal Node. The producers, who irrigated more than 7.5 times, yielded 5669 kg da^{-1}, and the producers, who irrigated 7.5 times or less, yielded 4,471 kg da^{-1}. It can be said that producers need to irrigate more than 7 times. Helyes et al. (2012) reported that irrigation has a greater effect on fruit weight compared to number of fruit. Also, irrigated plants showed significantly higher yields and rain fed plants lost yield. Node 2 was divided into two sub-nodes (Node 5-6), and Node 6 was identified as the second terminal Node. Node 6, in which producers with higher education level were, provided 8011 kg of product per decare, and moreover, the producers with the highest yield are in the Node 6 group. It can be stated that increasing the education level of the producers provides more conscious production and higher efficiency. Node 3 is divided into two sub-nodes, which are Node 7 (NS ≤ 1183) and Node 8 (NS > 1183). Node 5 is divided into two terminal Nodes which are Node 9 (NI≤9.5), in which yield is 5651 kg da^{-1}, and Node 10 (NI>9.5), in which yield is 7,414 kg da^{-1}. Node 8 is divided into 2 sub-nodes (Nodes 11-12) in terms of ACF. Tomato yield for Node 11, in which 74 kg da^{-1} or less fertilizer was applied, was found as 4,637 kg da^{-1}, whereas tomato yield for Node12, in which more than 74 kg da^{-1} fertilizer was applied, was found as 5,718 kg da^{-1}. Producers should apply more than 74 kg of fertilizer per decare. This observation is similar to the previous studies that have reported usage of fertilizers have a significant effect on the yield of vegetable crops (Haworth, 1961; Svec et al., 1976; Toor et al., 2006; Wang & Xing, 2017). The Node

11 is further divided into two sub-nodes, which are Nodes 13-14. Node 13 was characterized with AM ≤ 0.488 and yield was 4249 kg da^{-1}, whereas Node 14 was characterized with AM> 0.488 and yield was 4,961 kg da^{-1}. Besides, Nodes 12, 13 and 14 are terminal Nodes.

Figure 3 shows the decision tree diagram created by the CHAID algorithm. Node 0, which is TY, was divided two subgroups (Nodes 1-2). The number of seedlings was most affective factor on tomato yield.

Figure 2. The regression tree diagram created by the CART algorithm
Şekil 2. CART algoritması ile oluşturulan regresyon ağacı diyagramı

In the CHAID algorithm, NS, NI, SR, AF and EL independent variables were found to be significantly effective on TY (Adj. P. Value = 0.000, F = 18.601, df_1 = 2, df_2 = 102). In the CHAID decision tree diagram, the highest yield of 7,359 kg da^{-1} in Node 6 has been provided on the condition that NS is higher than 1,344 and SR is Karakoyunlu district.

It is understood that MARS data mining algorithm has higher prediction performance in TY estimation compared to other algorithms. In TY estimation, independent variables are included in the MARS estimation model given below.

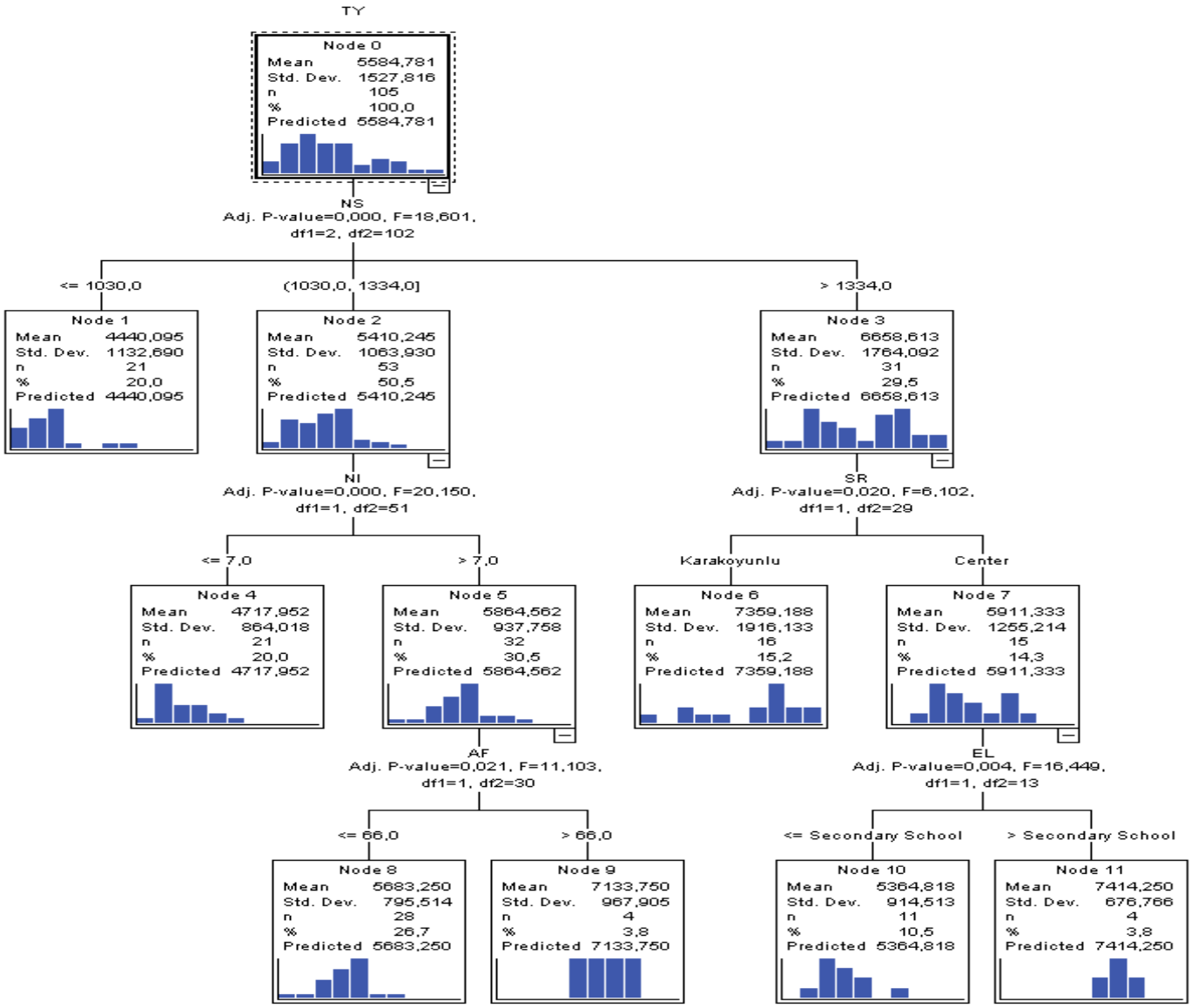

Figure 3. The regression tree diagram created by the CHAID algorithm
Şekil 3. CHAID algoritması ile oluşturulan regresyon ağacı diyagramı

TY = 4,735 + 43.95*max(0; NS-1450) – 4.15*max(0; 1,450-NS) + 378.75*max(0; NI-4) – 50.41*max(0; NS-1,450)*max(0; SAS_2) + 79.51*max(0; NI-4)*max(0; AF-62)*max(0; SR_2) + 0.5908*max(0; ACF-51)*max(0; NS-1,200) – 0.0258*max(0; ACF-51)*max(0; AF-27)*max(0; NS-1,200) + 1.85*max(0; ACF-51)*max(0; AF-48) + 0.6390*max(0; ACF-51)*max(0; 1,200-NS)*max(0; EL_5) + 317.02*max(0; NI-4)*max(0; EL_2)*max(0; SR_2) + 6.60*max(0; ACF-51)*max(0; 48-AF)*max(0; SAS_2) – 2.23*max(0; ACF-51)*max(0; NI-4)*max(0; 48-AF)

In the model; if NS <1450, it takes 43.95 * max (0; NS-1,450) = 0, and when NS> 1,450, it takes the value 43.95x NS-1,450. When random values are given to the independent variables that are determined to be significantly effective on TY in the MARS model, for example, as follows;

NI = 8, ACF = 80, AF = 42, NS = 1500, EL = 4, SAS = 1 and SR = 1;

TY = 4,735 + 43.95 * max (0; AFM-1,450) + 378.75 * max (0; NI-4) + 0.5908 * max (0; ACF-51) * max (0; AFM-1,200) - 0,0258 * max (0; ACF-51) * max (0; AF-27) * max (0; AFM-1,200) - 2.23 * max (0; ACF-51) * max (0; NI-4) * max (0; 48-AF)

TY = 4,735 + 43.95 * max (0; 1,500-1,450) + 378.75 * max (0; 8-4) + 0.5908 * max (0; 80-51) * max (0; 1,500-1,200) - 0.0258 * max (0; 80-51) * max (0; 42-27) * max (0; 1,500-1,200) - 2.23 * max (0; 80-51) * max (0; 8-4) * max (0; 48-42)

TY = 8731.12 kg da^{-1} (tomato yield can be obtained)

According to MARS model results, it is understood that the seedling number should be more than 1450, the number of irrigations should be more than 4, the cultivation area should be Karakoyun district and the fertilizer amount should exceed 51 kg da^{-1}. Due to the lack of similar statistical analysis methods in this field, the results of this study could not be compared to the literature. It is hoped that the MARS prediction model used in this study will contribute to the literature so that similar studies can be conducted in the future.

In this study, the comparison of tomato yield prediction powers of some data mining algorithms were conducted; as a result, some factors affecting tomato yield significantly were determined. NS, NI, EL, ACF and AM independent variables were determined to be statistically significant in CART algorithm, while NS, NI, SR, AF and EL independent variables were determined to be statistically significant in CHAID algorithm. Besides, NI, ACF, AF, NS, EL, SAS and SR are significant variables for MARS algorithm. The significance order of Pearson correlation coefficients between real and predicted values in tomato yield was determined as MARS (0.848a)> ANN (0.751b)> GLM (0.742)> CART (0.710)> CHAID (0.696b)> Exhaustive CHAID (0.682b). The MARS algorithm outperformed among the applied algorithms. Preferring MARS gives an opportunity to detect factors affecting tomato yield and their interactions. It was understood that the MARS algorithm may offer good solutions to farmers for making accurate decisions to increase tomato yield because of the fact that it is more informative with the best predictive accuracy. We hope that this study will contribute to paving the way for similar studies in the field of agriculture.

Researchers Contribution Rate Declaration Summary

The contribution of the authors is equal.

Conflict of Interest

The authors declare that there is no conflict of interest

REFERENCES

Anonymous, (2018). Food and Agricultural Commodities Production Database. http://faostat.fao.org/site/339/default.aspx (Date accessed: 12.05.2021).

Anonymous, (2019). Crop Production Statistics. https://www.tuik.gov.tr/Home/Index (Date accessed: 12.02.2021).

Anonymous, (2020). Temperature Data for the Province of Igdir. https://tr.climate-data.org/asya/tuerkiye/igd%C4%B1r-693/ (Date accessed: 12.03.2021).

Aytekin, İ., Eyduran, E., Karadaş, K., Akşahan, R., & Keskin, İ. (2018). Prediction of fattening final live weight from some body measurements and fattening period in young bulls of crossbred and exotic breeds using MARS data mining algorithm. *Revista Brasileira de Zootecnia 50*(1), 189-195. http://doi.org/10.17582/journal.pjz/2018.50.1.189.195

Bostancı, B. & Eren-Atay, C. (2018). Decision support tools for barley yield: the case of Menemen – Turkey. *Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering 20*(60), 1057-1067. https://doi.org/10.21205/deufmd.2018206085

Camdeviren, H.A., Yazici, A.C., Akkus, Z., Bugdayci, R., & Sungur, M.A. (2007). Comparison of logistic regression model and classification tree: an application to postpartum depression data. *Expert Systems with Applications 32*(4), 987–994. https://doi.org/10.1016/j.eswa.2006.02.022

Celik, S., Eyduran, E., Tatliyer, A., Karadas, K., Kara, M.K., & Waheed, A. (2018). comparing predictive performances of some nonlinear functions and multivariate adaptive regression splines (MARS) for describing the growth of daera dın panah (DDP) goat in Pakistan. *Pakistan Journal of Zoology 50*(3): 1-4. http://doi.org/10.17582/journal.pjz/2018.50.3.sc2

Cho, W., Na, M. & Park, Y. (2018). Extraction of optimum condition of cultivation factors to improve tomato production using statistical regression analysis and response surface methodology. *Advanced Science Letters 24*(3), 2084-2087.

Comlekcioglu, N. & Şimşek, M. (2014). The effect of gibberellic acid (GA3) on fruit set in industrial tomato at high temperature conditions and different water level. *Yuzuncu Yil University Journal of Agricultural Science 24*(3), 270- 279.

Debela, K. B., Belew, D., & Nego, J. (2016). Evaluation of tomato (Lycopersicon Esculentum Mill.) varieties for growth and seed quality under jimma condition, South Western Ethiopia. *International Journal of Crop Science and Technology 2*(2), 69-77.

Degefa, G., Benti, G., Jafar, M., Tadesse, F., & Berhanu, H. (2019). Effects of intra-row spacing and n fertilizer rates on yield and yield components of tomato (Lycopersicon Esculentum L.) at Harawe, Eastern Ethiopia, *Journal of Plant Sciences 7*(1), 8-12. https://doi.org/10.11648/j.jps.20190701.12

Duru, M., Duru, A., Karadas, K., Eyduran, E., Cinli, H., & Tariq, M.M. (2017). Effect of carrot (Daucus carota) leaf powder on external and internal egg characteristics of hy-line white laying hens. *Pakistan Journal of Zoology 49*(1), 125-132. https://doi.org/10.17582/journal.pjz/2017.49.1.125

Estrada-Ortiz, E., Trejo-Tellez, L.I., Gomez-Merino, F.C., Nunez-Escobar, R., & Sandoval-Villa, M. (2013). The effects of phosphite on tomato yield and fruit quality. *The Journal of Soil Science and Plant Nutrition 13*(3), 612–620. https://doi.org/10.4067/S0718-95162013005000049

Eyduran, E., Zaborski, D., Waheed, A., Celik, S., Karadas, K. & Grzesiak, W. (2017). Comparison of the predictive capabilities of several data mining algorithms and multiple linear regression in the prediction of body weight by means of body measurements in the indigenous beetal goat of Pakistan. *Pakistan Journal of Zoology 49*(1), 257-265. https://doi.org/10.17582/journal.pjz/2017.49.1.257.265

Friedman, J.H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* 19(1), 1-67. https://doi.org/10.1214/aos/1176347963

Hahn, F. (2013). Sensing, control and instrumentation during tomato growth. tomatoes: cultivation, varieties and nutrition. Food science and technology. Nova Science Publishers, 339pp.

Haworth, F. (1961). The effects of organic and inorganic nitrogen fertilisers on the yield of early potatoes, spring cabbage, leeks and summer cabbage. *Journal of Horticultural Science 36*, 202-205

Helyes, L., Lugasi, A., & Pek, Z. (2012). Effect of irrigation on processing tomato yield and antioxidant components. *The Turkish Journal of Agriculture and Forestry 36*(6), 702-709. https://doi.org/10.3906/tar-1107-9

Irmak, S., & Ercan, U. (2017). Determining of the affecting factors edible oil consumption using data mining method. *Kafkas University Economics and Administrative Sciences Faculty the Journal 8*(15), 57-79.

Karadas, K. & Birinci, A. (2019). Determination of factors affecting dairy cattle: a case study of ardahan province using data mining algorithms. *Revista Brasileira de Zootecnia 48*: 1-11. https://doi.org/10.1590/rbz4820170263

Karadas, K. & Kadirhanogullari, I.H. (2017). Predicting honey production using data mining and artificial neural network algorithms in apiculture. *Pakistan Journal of Zoology 49*(5), 1611-1619. https://doi.org/10.0.68.174/journal.pjz/2017.49.5.1611.1619

Kibria, G., Islam, M., & Alamgir, M. (2016). Yield and nutritional quality of tomato as affected by chemical fertilizer and biogas plant residues. *International Journal of Plant & Soil Science 13*(2), 1-10. https://doi.org/10.9734/IJPSS/2016/29434

Kiracı, S. & Karataş, A. (2015). Organic tomato growing plant activator applications effects on yield and quality. *Journal of Adnan Menderes University Agricultural Faculty 12*(1), 17–22.

Küçükönder, H., Vursavuş, K.K., & Üçkardeş, F. (2015). Determining the effect of some mechanical properties on color maturity of tomato with K-Star, Random Forest and Decision Tree (C4.5) Classification Algorithms. *Turkish Journal of Agriculture - Food Science and Technology 3*(5), 300-306. https://doi.org/10.7161/omuanajas.952786

Letourneau, G., Caron, J., Anderson, L., & Cormier, J. (2015). Matric potential-based irrigation management of field-grown tomato: effects on yield and water use efficiency. *Agricultural Water Management 161,* 102–113. https://doi.org/10.1016/j.agwat.2015.07.005

Liu, J., Hu, T., Feng, P., Wang, L., & Yang, S. (2019). Tomato yield and water use efficiency change with various soil moisture and potassium levels during different growth stages. *Plos One 14*(3), 1-14. https://doi.org/10.1371/journal. pone.0213643.

Manan, A., Ayyub, C.M., Aslam, Pervez, M., & Ahmad, R. (2016). Methyl jasmonate brings about resistance against salinity stressed tomato plants by altering biochemical and physiological processes. *The Pakistan Journal of Agricultural Sciences 53*(1), 35-41. https://doi.org/10.21162/PAKJAS/16.4441

Na, M., Park, Y., & Cho, W. (2017). A study on optimal environmental factors of tomato using smart farm data. *Journal of the Korean Data & Information Science Society 28*(6), 1427–1435. https://doi.org/10.7465/jkdi.2017.28.6.1427

Navarro-Gonzále, I., García-Alonso, J., & Periago, M. J. (2018). Bioactive compounds of tomato: Cancer chemopreventive effects and influence on the transcriptome in hepatocytes. *Journal of Functional Foods 42*, 271-280. https://doi.org/10.1016/j.jff.2018.01.003

Neta, M.N.A., Mota, W.F., Pegoraro, R.F., Pacheco, M.C., Batista, C.M., & Sorases, M.C. (2019). Agronomic yield and quality of industrial tomatoes under NPK doses. *Revista Brasileira de Engenharia Agrícola e Ambiental 24*(1), 59-64. https://doi.org/10.1590/1807-1929/agriambi.v24n1p59-64

Özer, H. (2016). Organic tomato production. international journal of agricultural and wildlife sciences, *Abant Izzet Baysal University Faculty of Agriculture and Natural Sciences 2*(1), 43–53.

Özkan, Z., Ünlü, L., & Ögür, E. (2017). Comparison of the efficiency of pheromone and pherolite traps used against tomato moth (tuta absoluta meyrick) in greenhouse tomato growing. *Harran Journal of Agricultural and Food Sciences 21*(4), 394-403. https://doi.org/10.29050/harranziraat.290747

Sajjad, M., Ashfaq, M., Suhail, A., & Akhtar, S. (2011). Screening of tomato genotypes for resistance to tomato fruit borer (helicoverpa armiger hubner) in Pakistan. *The Pakistan Journal of Agricultural Sciences 48*(1), 59-62.

Söylemez, S., & Pakyürek, A. Y. (2017). Effect of different rootstocks and nutrient induced ec levels on element content of the tomato fruits. *Turkish Journal of Agricultural and Natural Sciences 4*(2), 155–161.

Sun, J. & Hui, L. (2008). Data mining method for listed companies, financial distress prediction. *Knowledge-Based Systems 21*(1), 1-5. https://10.1016/j.knosys.2006.11.003

Svec, L. V., Thoroughgood, C. A., & Mok, H. C. S. (1976). Chemical evaluation of vegetables grown with conventional or organic soil ammendments. *Communications in Soil Science and Plant Analysis 7*(2), 213-228. https://doi.org/10.1080/00103627609366634

Tapiero, H., Townsend, D. M., & Tew, K. D. (2004). The role of carotenoids in the prevention of human pathologies. *Biomedicine and Pharmacotherapy 58*(2): 100-110. https://doi.org/10.1016/j.biopha.2003.12.006

Tatar, M., & Pirinç, V. (2017). Potential of industrial tomato production of Southeast Anatolian Region in Turkey. *Iğdır University Journal of the Institute of Science and Technology 7*(2), 11-20. https://doi.org/10.21597/jist.2017.121

Tesafay, T., Gebremariam, M., Gebredsadik, K., Hagazi, M., & Girmay, S. (2018). Tomato yield and economic performance under vermicompost and mineral fertilizer applications. *The Open Agriculture Journal 12*(1), 262-269. https://doi.org/10.2174/ 1874331501812010262

Toor, R.K., Geoffrey, P.S., & Anuschka, H. (2006). Influence of different types of fertilisers on the major antioxidant components of tomatoes. *Journal of Food Composition and Analysis 19*(1), 20–27. https://doi.org/10.1016/j.jfca.2005.03.003

Wang, X., & Xing, Y. (2017). Evaluation of the effects of irrigation and fertilization on tomato fruit yield and quality: a principal component analysis. *Scientific reports 7*(1), 350. https://doi.org/10.1038/s41598-017-00373-8

Yamane, T. (2001). *Turkish Translation of the Basic Sampling Methods*. Translators: Esin, A., Aydın, C., Bakır, M.A., Gürbüzsel, E., Literatür Yayınları, Tukey. pp.509.

Yaraş, G., & Daşgan, H. Y. (2012). Effects of soil-applied micronized-sulphur with bentonite and organic matter on soil ph, tomato plant growth, yield and fruit quality under greenhouse conditions. *Reserach Journal of Agricultural Sciences 5*(1): 175-180.