EURASIAN JOURNAL OF FOOD SCIENCE AND TECHNOLOGY

Optimization Using Decision Trees Method in Multivariable Food Engineering Experiments and Its Sample of Applicability on Experiment Related with the Nisin Production of *Lactococcus lactis* N8

Selahaddin Batuhan AKBEN

Osmaniye Korkut Ata University, Bahce Vocational School, Department of Computer Technologies, Osmaniye/Turkey

*Corresponding Author: batuhanakben@osmaniye.edu.tr

Abstract

In this study, the ranges of independent variables resulting the optimum result of experiment should be selected was determined using decision trees method. Thus, the applicability of decision trees method has been proposed to food engineering experiments aiming the optimization. The sample application of the decision tree method proposed in the study was performed in the experiment aiming optimum nisin production of *Lactococcus lactis* N8. According to the findings obtained from the sample application it was observed that the decision trees method determines both optimum variable values and their tolerance ranges. Furthermore, the method proposed was not only determined the optimal ranges of variable values also it was determined the variable ranges for all possible experimental results. Accordingly, at the end of the study, advantages of the proposed method were explained by comparing with similar methods and how the experimental design should be to make the method more effective was proposed.

Keywords: Decision Trees, Modelling, Nisin Production, Optimization, Prediction

INTRODUCTION

In the field of food engineering, experiments are carried out for optimization especially in microbiology and biotechnology experiments (Mandenius et al., 2008; Kalkan et al., 2014). In these optimization experiments, experimental results are investigated depend on the interaction of independent variables and their variation of the variable values. Thus, it is determined what the variable values that maximize or minimize the experiment result should be (Banga, et al., 2003). Many methods especially dependent on curve fitting are used as optimization methods (Kalkan et al., 2016; Saguy et al., 1984). The well-known and most used of these methods is the response surface methodology that is depending on second-order curve fitting. However, even in this method, only the coefficients that produce the optimum result can be determined. Namely the optimal ranges of independent variables that produce the requested result are not provided precisely in the response surface methodology. For this reason, it is not much sensitive to the external factors can change the experimental results (Myres et al, 1995; Baş et al, 2007).

Moreover, if the optimum experimental results are more than one, the current methods can determine only one of them. Therefore, it cannot determine also the variable values that will create the second optimal experimental result. Apart from all these, the number of optimization methods used in the field of food engineering is not much more (Koç et al., 2010).

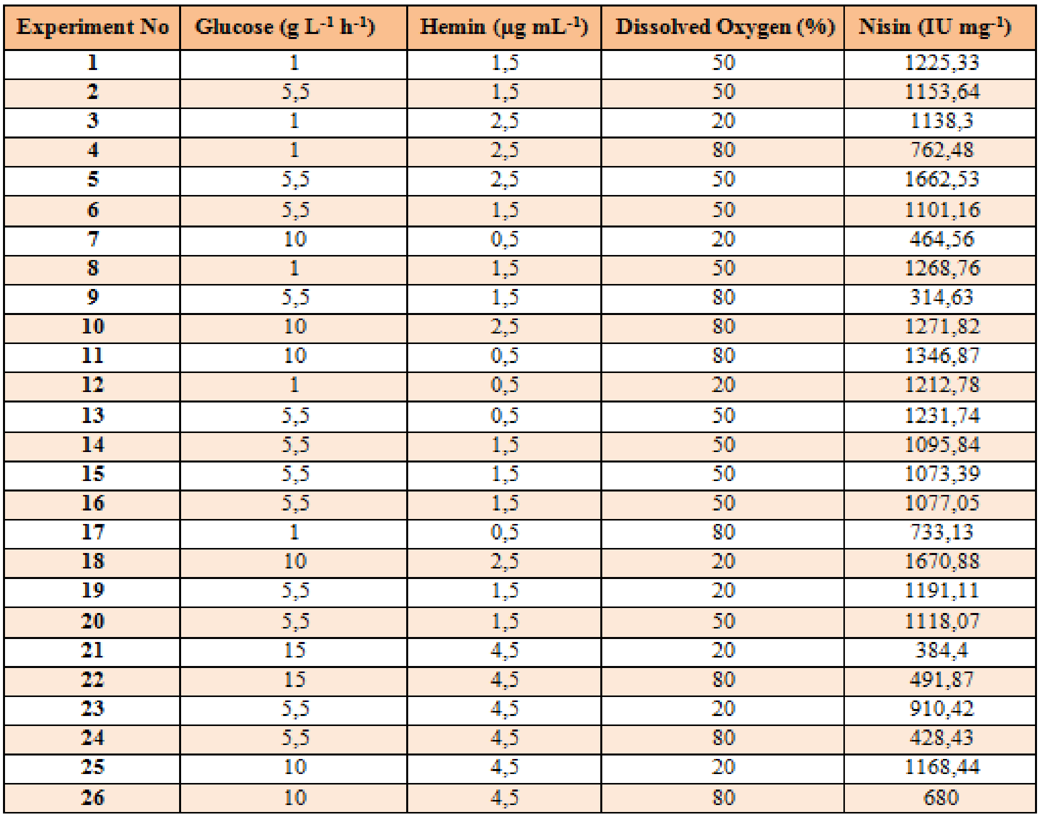
For this reason, a method that will provide the optimum ranges of independent variable values to produce requested experimental results has been proposed as alternative optimization method in this study. This method proposed was applied to the sample food engineering experiment and tested. According to the findings obtained, advantages were also determined.

MATERIAL and METHOD

Material

In this study, the sample experiment is the nisin production of Lactococcus lactis N8 with hemin-stimulated cell respiration in the fed-batch fermentation system. This experiment was previously performed using response surface methodology and is in the literature (Kördikanlıoğlu, 2014; Kördikanlıoğlu et al, 2015). Thus, the comparison of the proposed method with the surface response methodology can be done by using the data of this experiment. Also, since the same experiment is in the literature, details of the experiment are not given in this study. Variables are glucose, hemin and oxygen concentrations used in fed-batch fermentation. Tried variable values and the results obtained are also shown in the Table 1.

Table 1. Variable value variations used in the sample experiment and the results obtained (Experimental Design used for Optimization of Hemin, Glucose and Dissolved Oxygen Concentration in the Fed-batch Fermentation System)



Experiment No	Glucose (g L^{-1} h^{-1})	Hemin (µg mL^{-1})	Dissolved Oxygen (%)	Nisin (IU mg^{-1})
1	1	1,5	50	1225,33
2	5,5	1,5	50	1153,64
3	1	2,5	20	1138,3
4	1	2,5	80	762,48
5	5,5	2,5	50	1662,53
6	5,5	1,5	50	1101,16
7	10	0,5	20	464,56
8	1	1,5	50	1268,76
9	5,5	1,5	80	314,63
10	10	2,5	80	1271,82
11	10	0,5	80	1346,87
12	1	0,5	20	1212,78
13	5,5	0,5	50	1231,74
14	5,5	1,5	50	1095,84
15	5,5	1,5	50	1073,39
16	5,5	1,5	50	1077,05
17	1	0,5	80	733,13
18	10	2,5	20	1670,88
19	5,5	1,5	20	1191,11
20	5,5	1,5	50	1118,07
21	15	4,5	20	384,4
22	15	4,5	80	491,87
23	5,5	4,5	20	910,42
24	5,5	4,5	80	428,43
25	10	4,5	20	1168,44
26	10	4,5	80	680

Method

Decision Trees Method

Decision trees are a decision support system that is used to determine which variable value variations produce which result values. Decision trees achieve this goal using different mathematical algorithms. In this study, ID3 algorithm based on entropy and information gain is used. This algorithm primarily measures the information gain between the variables and experimental result. Then the variable with highest information gain is divided into clusters containing the same values. Then the same procedure is applied to the remaining variables and subsets are generated. Also determines the variable limits that must be selected to create each of subsets during clustering (Rokach et al., 2014; Ville., 2006; Akben et al., 2016) The information gain between any variable and corresponding result is as in Equation 1. In Equation 1, x is a variable, x_i is the i^{th} value of variable x then z is the experimental result.

$$I(x) = E_z - \sum_{i=1}^{i=n} P(x_i, x) E_{x_i} \tag{1}$$

In the Equation 1, E is the entropy value as in Equation 2 and P is the probability function as in Equation 3.

$$E_{x_i} = -\sum_{c=1}^{c=m} P\left(z_c, \sum z_c\right) \, log_2{}^{P(z_c, \sum z_c)} \tag{2}$$

In the Equation 2, m is the number of result values corresponding to the variable x_i while result values corresponding to the variable x_i are z_c.

$$P(x_i, x) = \frac{\sum x_i}{\sum x} \tag{3}$$

In Equation 3, the sum of each x_i value is divided by the sum of all x values. The calculation of the entropy value for a variable value on the sample experimental design is as in Fig.

$$E_{x_1} = -1/3\log_2{}^{1/3} - 2/3\log_2{}^{2/3}$$

$$E_Z = -1/5\log_2{}^{1/5} - 2/5\log_2{}^{2/5} - 2/5\log_2{}^{2/5}$$

$$I(x) = E_Z - 3/5E_{x_1} - 2/5E_{x_2}$$

$$\begin{bmatrix} x_1 & y_1 \\ x_1 & y_2 \\ x_1 & y_2 \\ x_2 & y_2 \\ x_2 & y_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_2 \\ z_3 \\ z_3 \end{bmatrix}$$

x_1 is the first value of variable x
x_2 is the second value of variable x

y_1 is the first value of variable y
y_2 is the second values of variable y

z_1 is the first value of experimental result
z_2 is the second value of experimental result
z_3 is the third value of experimental result

Figure 1. Calculation of the entropy value for a variable value on the sample experimental design

In Figure 1, the decision tree result generated in single process for two variables. However, in this study the decision tree was applied separately for each variable since the experimental model results corresponding to the variables were not equal. Thus, the optimal limits for each variable were determined separately.

Curve Fitting Method based on Vandermonde matrix

The polynomial model obtained by curve fitting was used to evaluate the effect of modeling on decision tree use in this study. The polynomial model is based on the Vandermonde matrix. Thus, the polynomial model used can also be called the Vandermonde polynomial (Cazals et al., 2005). The method of creating Vandermonde polynomials is as follows.

Assume that $\mathrm{V}_{\mathrm{n,m}}$ is the n^{th}value of m^{th}variable and $\mathrm{R}_{\mathrm{n,m}}$ is the experimental result values corresponding to this variable. In this case, the relation between the $\mathrm{V}_{\mathrm{n,m}}$ and $\mathrm{R}_{\mathrm{n,m}}$ is as in Equation 1. Thus, the polynomial model can be represented as is in Equation 2.

$$V_{n,m}{}^{n} + \cdots + c_2 V_{n,m}{}^{2} + c_1 V_{n,m}{}^{1} + c_0 V_{n,m}{}^{0} = R_{n,m} \qquad (1)$$

The c_{n} values in Equation 1 are polynomial coefficients can be calculated as in Equation 2.

$$\begin{bmatrix} V_{0,m}{}^{0} & V_{0,m}{}^{1} & V_{0,m}{}^{2} & \cdots & V_{0,m}{}^{n} \\ V_{1,m}{}^{0} & V_{1,m}{}^{1} & V_{1,m}{}^{2} & \cdots & V_{1,m}{}^{n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ V_{n,m}{}^{0} & V_{n,m}{}^{1} & V_{n,m}{}^{2} & \cdots & V_{n,m}{}^{n} \end{bmatrix} \begin{bmatrix} c_{0,m} \\ c_{1,m} \\ \vdots \\ c_{n,m} \end{bmatrix} = \begin{bmatrix} R_{0,m} \\ R_{1,m} \\ \vdots \\ R_{n,m} \end{bmatrix} \qquad (2)$$

In this study, each polynomial equation represents relation between the experimental result and one of variables. For this reason, a different polynomial equation was created for each variable in the study.

RESULTS and DISCUSSION

Decision trees method was tested for both raw variable values and polynomial model in this study. Thus, it is also determined whether modeling is necessary if decision trees method will use. Experiment result values for the test of proposed method were divided into 4 classes for each 500 IU mg-1 interval. That is, each of result values in the range of 0-500, 500-1000-1500, 1500 < IU mg-1 was considered as a single value and experimental result values were assigned as 4 class. This process allows the creation of limit values for each interval. The optimum limits determined by decision trees for both raw and modeled variable values are the as in the Table 2.

Table 2. Optimum limits for both raw and modelled variable values

For Nisin > 1500 IU mg^{-1}	Glucose (g L^{-1} h^{-1})	Hemin (µg mL^{-1})	Dissolved Oxygen (%)
Limits for raw data	">3.25" *and* " < 15"	" > 2" *and* " < 3.5"	" > 20" *and* " < 65"
Optimal values of raw data (Mean of Limits)	9.125	2.75	42.5
Limits for modelled data	">5.83" *and* " < 11.57"	" > 2.36" *and* " < 3.64"	" > 31.1" *and* " < 56.3"
Optimal values of modelled data (Mean of Limits)	8.7	3	43.7

As it is seen in the table, applying the decision trees to experimental data obtained by modeling was produced more applicable (narrower) limit ranges as compared to raw experimental data. If so, it would be more appropriate to apply the decision trees to the experimental results obtained by modeling. The modeling applied in the study is to create the Vandermonde polynomials. Number of the variables values tried in the study was 4 for hemin and glucose and 3 for dissolved oxygen rate. Therefore, the polynomial degree for the hemin and glucose was selected as 3 and for the dissolved oxygen rate was selected as 2. The experimental values produced by the model were normalized to reduce the error rate. That is, in accordance with the actual distribution ratio the calculated result values were placed (stretched) between the largest and smallest experimental result of the raw data. The normalized graphs obtained by polynomial models can be seen from Figure 2.

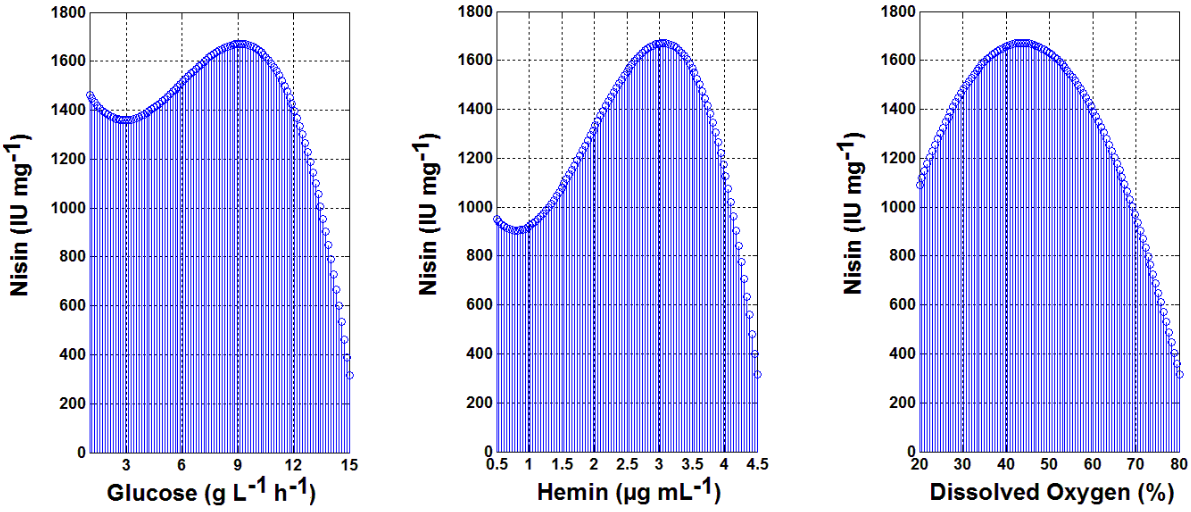


Figure 2. Graphs that obtained by normalized polynomial models.

The limits of optimal variable values calculated by decision trees can be seen in Figure 3. The cyan color cube in Figure 3 is the space that results of the experiment will be nearly optimized while the red dot is precise coordinate of optimal experimental result. Also, the red dot is the center of the cyan color cube.

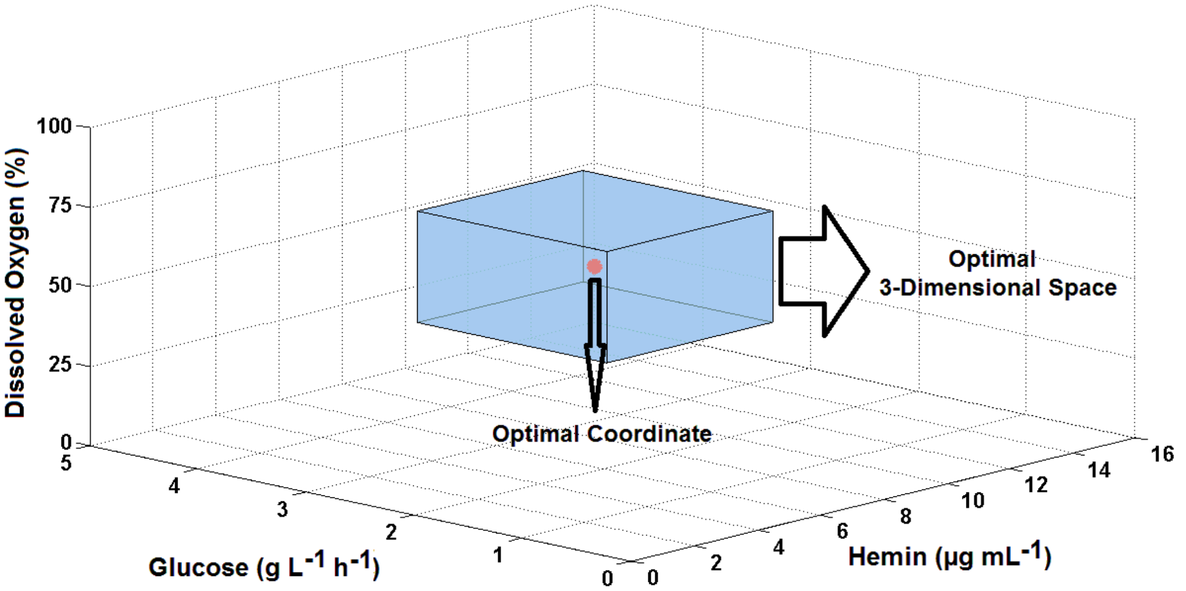


Figure 3. The tolerance limits of optimal variable values and the precise optimal coordinate that were created by decision trees to produce the optimal experimental results

These limits in Figure 3 are also related to the graphical results generated by modeling in Figure 2. In other words, the limits of optimal values created by decision trees are the same in both Figures 2 and 3. The Figure 2, Figure 3 and Table 2 show that the decision trees can very successfully determine the optimal variable values that produce the optimal experimental result because the optimal variable values determined by response surface method in literature are nearly same.

In the same optimization experiment in which the response surface analysis method was used in the previous study, it was determined that optimal values were 3 for glucose, 8 for hemin and 40 for dissolved oxygen. These previous study results are similar to the results of this study while determined optimal oxygen values are slightly different. So the decision trees method can be reached to goal of the response surface methodology. However, decision tree method has some advantages as compared to response surface methodology. One of these advantages is that the decision trees method not only determines the optimal test result but provides the interval (Cyan cube in the Figure 3) of optimal experimental results also. This makes the optimal coordinate more sensitive to possible changes that may occur due to various factors.

Another advantage is that decision trees can determine the limits of variable values for all experimental results. So that in experiments if there is two optimal experimental results the optimal variable value ranges can be determined for both coordinates. Table 3 shows the optimal range of variable values for all experimental result limits determined by decision trees method.

Table 3. Optimum limits of variable values to be selected to produce the all experimental result ranges

Nisin (IU mg^{-1})	**Glucose (g L^{-1} h^{-1})**	**Hemin (µg mL^{-1})**	**Dissolved Oxygen (%)**
For Nisin<500	" > 14.65" *and* " < 15"	" > 4.4" *and* " < 4.5"	" > 77.3" *and* " < 80"
For 1000>Nisin>500	" > 13.53" *and* " < 14.65"	> 0.5 *and* <1.32" *or* " > 4.12 *and* < 4.4"	" > 69.5" *and* " < 77.3"
For 1500>Nisin>1000	" > 1" *and* "<5.83" *or* " > 11.57" *and* " < 13.53"	" > 1.32 *and* <2.36" *or* " > 3.64 *and* < 4.12"	" > 20" *and* "<31.1" *or* " > 56.3" *and* " < 69.5"
For Nisin>1500	">5.83" *and* " < 11.57"	" > 2.36" *and* " < 3.64"	" > 31.1" *and* " < 56.3"

As seen in the Table 3, decision trees method can calculate limits for the experimental design and determine relation between the experimental results and these limits. If desired, more precise limits can be specified by reducing the ranges of experimental results.

Furthermore, the decision trees method is compatible with all modeling algorithms. Even, it can also achieve the goal using the raw experimental data without modeling. In this case, the decision trees method eliminates one of the response surface methodology disadvantage arising from second degree polynomial dependency. In addition, the experimental design should be performed with a linear increase of the variable values so that the proposed method can determine the variable value intervals more efficiently. Because, if the variable values are not homogenous in the experimental design, the optimal variable values determine in very wide ranges and error possibility increase.

CONCLUSION

In this study, the use of decision tree method in food processing experiments has been proposed. The proposed method has been tested on a sample experiment the result is dependent on 3 variables. According to the findings obtained, the method has not only determined the variable values to produce the optimal experimental result but also determined the tolerance ranges of optimal variable values. Thus, the ranges have also been determined for experiments that the optimal variable values may possibly change depending on various factors. In addition, the proposed method can also determine the both optimal variable value ranges in cases where there is more than one optimal experimental result in the experiments. Furthermore, the proposed method can be used together with all modeling methods as well as being able to achieve its purpose without modeling also. As a result, it can be said that proposed method can eliminate the many deficiencies of the response surface methodology which is the well-known and mostly used current optimization method of food processing experiments. Moreover, the proposed method aiming at optimization will be an alternative to the field of food engineering which is quite lacking in terms of optimization methods.

ACKNOWLEDGMENTS

This article was presented at the International Conference on Agriculture, Forest, Food Sciences and Technologies(ICAFOF) conference held in Cappadocia / Nevşehir on May 15-17, 2017 and published in summary.

The sample experiment in which the method proposed of this study was tested on is in the literature. I thank to Burcu Özel (She prepared and used the same experiment in her article and Msc thesis available in the literature) for her help in understanding the experiment.

REFERENCES

Akben S. B., Alkan A. 2016. Analysis of Yatch Hydrodynamics Scale-Effect and Optimal Scale Range Determination by Using Decision Tree Algorithm, *International Conference on Natural Science and Engineering (ICNASE' 16), Kilis/Turkey*, 412-418.

Banga J. R., Balsa-Canto E., Moles C. G., Alonso A. A. 2003. Improving food processing using modern optimization methods, *Trends Food Sci & Tech*, 14, 131-144.

Baş D., Boyacı İ. H. 2007. Modeling and optimization I: Usability of response surface methodology, *J Food Eng.*, 78, 836-845.

Cazals F., Pouget M. 2005. *Estimating differential quantities using polynomial fitting of osculating jets*, Computer Aided Geometric Design, 22(2), 121-146.

Kalkan S., Ünal E., Erginkaya Z. 2014 Nisin ilave edilmiş metil selüloz filmlerin antimikrobiyel etkilerinin belirlenmesi, *Gıda ve Yem Bilimi-Teknolojisi Dergisi/Journal of Food and Feed Science-Technology* 14, 1-7, 2014.

Kalkan S. 2016. Probiyotik Laktik Asit Bakterilerinin Staphylococcus aureus'a Karşı Antimikrobiyel Etkilerinin Farklı Matematiksel Modeller ile Analizi, *Sinop Üniversitesi Fen Bilimleri Dergisi*, 1(2), 150-159.

Kördikanlıoğlu“B. 2014. *Lactococcus lactis'te solunumun hemin ile teşvik edildiği yarı-kesikli fermentasyon sisteminde nisin üretiminin optimizasyonu*, Msc. thesis, Pamukkale University, Institute of Science, Denizli, Turkey.

Kördikanlıoğlu B. Şimşek Ö, Saris P. E. J. 2015. Nisin production of Lactococcus lactis N8 with hemin-stimulated cell respiration in fed-batch fermentation system, *Biotechnology Progress*, 31(3), 678-685.

Mandenius C. F., Brundin A. 2008. Bioprocess optimization using design of experiments methodology, *Biotechnology progress*, 24(6), 1191-1203.

Myers R. H., Montgomery D. C. 1995. *Response Surface Methodology, Process and Product Optimization Using Designed Experiments, 2nd ed.* John Wiley and Sons, New York.

Koç B., Ertekin F. K. 2010. Yanıt Yüzey Yöntemi ve Gıda İşleme Uygulamaları, *GIDA/The Journal of Food*, 35(1), 1-8.

Rokach L, Maimon O. 2014. *Data mining with decision trees: theory and applications*, World scientific.

Saguy I., Mishkin M. A., Karel M. 1984. Optimization methods and available software, part1, *CRC Critical RevFood Sci and Nutr*, 20(4), 275-299.

Ville B. D. 2006. *Decision tree for business intelligence and data mining*, SAS Publishing.