Gaziosmanpaşa Üniversitesi Ziraat Fakültesi Dergisi
Journal of Agricultural Faculty of Gaziosmanpasa University
http://ziraatdergi.gop.edu.tr/

JAFAG
ISSN: 1300-2910
E-ISSN: 2147-8848
(2020) 37 (1), 9-16
doi:10.13002/jafag4646

Araştırma Makalesi/*Research Article*

Classification of Apple Varieties: Comparison of Ensemble Learning and Naive Bayes Algorithms in H2O Framework

Dilara GERDAN[1], Abdullah BEYAZ[1*], Mustafa VATANDAŞ[1]
[1]Ankara University, Faculty of Agriculture, Department of Agricultural Machinery and Technologies Engineering
(orcid.org/0000-0002-2705-299X); (orcid.org/0000-0002-7329-1318) ; (orcid.org/0000-0003-3235-042X)
* e-mail: abeyaz@ankara.edu.tr

Alındığı tarih (Received): 16.09.2019 Kabul tarihi (Accepted): 30.03.2020
Online Baskı tarihi (Printed Online): 14.04.2020 Yazılı baskı tarihi (Printed): 30.04.2020

Abstract: In this study, H2O machine learning classification techniques were used to classify the apples according to the skin color of the fruits. For each variety, 60 samples were used at evaluations of the fruits. Fruit color values were based on L *, a * and b * color space, and measured by a portable spectrophotometer. Red Delicious, Golden Delicious, and Granny Smith apple varieties were studied to create the database, randomly. H2O Gradient Boosting Machine, H2O Random Forest, and H2O Naive Bayes Algorithms were used for data analysis. The data set was partitioned to 30% for testing and 70% for training. The classifier performance which accuracy (%), error percentage (%), F-Measure, Cohen's Kappa, recall, precision, true positive (TP), false positive (FP), true negative (TN), false negative (FN) values were given at the conclusion section of the research. The results found that 100,0 % accuracy for H2O Gradient Boosting Machine, 98,4 % accuracy for H2O Random Forest and 100,0 % accuracy for H2O Naive Bayes.

Keywords: Apple classification, H2O machine learning, Gradient Boosting Machine, Random Forest, Naive Bayes

Elma Çeşitlerinin Sınıflandırılması: H2O Tabanlı Kollektif Öğrenme ve Naive Bayes Algoritmalarının Karşılaştırılması

Öz: Bu çalışmada, H2O tabanlı makine öğrenmesi sınıflandırma teknikleri, meyveleri kabuk rengine göre sınıflandırmak amacıyla kullanılmıştır. Veri seti oluşturmak için rastgele seçilen 60 adet Red Delicious, 60 adet Golden Delicious ve 60 adet Granny Smith elma çeşidine ait veriler değerlendirilmiştir. Meyve renk değerlerinde, L *, a * ve b * renk uzayı esas alınmış ve taşınabilir spektrofotometre ile ölçümler yapılmıştır. Veri analizi için H2O Gradyan Artırma Makinesi, H2O Rastgele Orman ile H2O Naive Bayes algoritmaları seçilmiştir. Veri seti test için %30, eğitim için ise %70 olarak bölümlendirilmiştir. Değerlendirme; doğruluk (%), yüzde hata (%), F-Ölçümü, Cohen's Kappa, hatırlama, doğruluk, doğru pozitif (TP), yanlış pozitif (FP), gerçek negatif (TN), yanlış negatif (FN) değerleri gibi performans göstergelerine göre yapılmıştır. Sonuçlar, H2O Gradyan Artırma Makinesi için %100,0, H2O Rastgele Orman için %98,4 ve H2O Naive Bayes algoritması için %100,0 doğrulukta elde edilmiştir.

Anahtar Kelimeler: Elma sınıflama, H2O makine öğrenmesi, Gradyan Artırma Makinesi, Rastgele Orman, Naive Bayes

1. Introduction

In today's economic life, it is the main objective of the economy to make maximum use of limited production factors. With Industry 4.0, it is foreseen that the highest income and yield will be achieved in the current conditions in agriculture. Industry 4.0 elements are the simulation, Internet of Things, Big Data, vertical and horizontal system organization, M2M (Machine -to- Machine) technology, cybersecurity, cloud computing, etc., have affected improved agricultural sector just as all other areas. In modern agriculture systems, mechanization and automation have gained importance in all stages, from production to consumption. Especially, harvested fruits and vegetables are served to domestic and foreign markets. Various technologies have been developed in this field. These technologies are mostly based on basic methods of cleaning and

classification. The cleaning process is intended to remove all foreign matter from the products. In the classification process, classification is made according to various characteristics. For this, the physical properties of the biological material, which is often utilized, can be seen in Table 1. The classification process is carried out to separate the products cleaned from foreign substances according to their types, size characteristics, and quality. Classification of biological material can be made according to physical, chemical and biological properties. In the classification process, according to the physical properties of biological materials could be separated based on; mechanical, thermal, optical and electrical properties. In chemical classification process, acid amount content, sugar amount content, tannin amount content, carbon dioxide amount content and pH values are calculated. Biological classification processes, also based on the degree of maturation, respiration, odor, taste, behavioral properties against biochemical substances, are examined (Mohsenin 1980; Öztürk 1988).

Table 1. Physical properties of biological material (Mohsenin 1980)
Tablo 1. *Biyolojik materyalin fiziksel özellikleri (Mohsenin 1980)*

Physical Properties			
Mechanical Properties	**Thermal Properties**	**Optical Properties**	**Electrical Properties**
Main dimensions	Specific heat	Color	Electrical conductance and capacitance
Geometrical dimensions	Thermal conductivity	Light reflectance and transmittance	Dielectric properties
Mass	Thermal diffusivity		Reaction to electromagnetic radiation
Density	Surface conductance		
Hardness	Emissivity		
Static and sliding coefficient			
Coefficient of friction			
Compressive strength			
Impact and shear resistance			

Ensemble methods (bagging, boosting, etc.) aim to improve the performance of a particular statistical learning or model technique. The general principle of community methods is to create a combination of some methods rather than using a single fit of the method (Bühlmann 2012). A community (ensemble) is a collection of estimators (average of all estimates) that come together to give a final estimate. The reason the use of the ensemble methods is that many different determinants who try to predict the same target variable will do a better job than any single predictor alone. Ensemble techniques are classified as Boosting, Bagging, AdaBoost, Stacking (blending, MAVL) (Şeker and Erdoğan 2018). Different machine learning applications are based on nonparametric regression and classification models of obtained data. A specific model designing can be done by using classical theories and adjusting some parameters according to the data properties (Natekin and Knoll 2013).

Classification is one of the main issues in machine learning and data mining. The aim of classification is predicting a set of training examples with class labels. For example, Bhatt et al. (2014) developed an apple classifier based on Artificial Neural Network (ANN). The authors said that a low level of error prediction confirmed the fact that the Neural Network model is an effective instrument of the apple quality estimation. Nandi et al. (2014) studied a computer vision-based system for mango fruit grading with Support Vector Machine (SVM). The performance of this system was 90% accuracy. Semary et al. (2014) developed a new

classification system for tomato-based on color and texture features. 177 tomato fruits each was captured from four sides, and data partitioned as 70% of the total images for the training phase and 30% for testing. Their proposed system achieved 92% accuracy. Zawbaa et al. (2014) worked on apples, strawberry, and oranges automatic classification based on the Random Forest (RF) algorithm. Experiments were tested using 178 fruit images. They imply that the Random Forest (RF) based algorithm provides better accuracy compared to the other well know machine learning techniques. Sofu et al. (2016) worked on automatic apple sorting based on real-time processing with the C4.5 algorithm. They imply that the C4.5 algorithm very fast and simple for sort of the apples. Canizo et al. (2019) created classification models that could predict wine origin information and compared them with data mining algorithms. Multiple Linear Regression (MLR), K Nearest Neighbor (k-NN), Support Vector Machines (SVM) and Random Forest (RF) algorithms were used to determine the 29 elements in grapes. The best results were estimated with SVM and RF algorithms with 84% and 88.9% accuracy, respectively.

2. Material and Methods

2.1. Data Collection

Apple varieties (Red Delicious, Golden Delicious, Granny Smith) were selected for this study, can be seen in Figure 1. 60 fruit were randomly chosen for each apple variety, totally of 180 fruit measured. After three readings, the average of the three data accepted as the main color value for each fruit. Totally 1620 measurement was done for creating the database. L*, a*, and b* values of fruits were measured by using the X-Rite Ci60 portable spectrophotometer (Figure 2). Three properties of 540 objects was created a database coming together.

Figure 1. Apple varieties (Red Delicious, Granny Smith, Golden Delicious, respectively)
Şekil 1. *Elma çeşitleri (sırasıyla Red Delicious, Granny Smith, Golden Delicious)*

2.2. H2O Framework

H2O is so fast, scalable, open-source machine learning and deep learning method for different applications (Aiello et al. 2016). H2O platform contains most of the Machine Learning (ML). It has some engines for parallel processing, analytics, and deep learning to have ML libraries (Suleiman and Al-Naymat 2017). H2O includes many common machine learning algorithms, such as generalized linear modeling (linear regression, logistic regression, etc.), Naive Bayes, principal component analysis, k-means clustering. H2O implements the best classification algorithms on scales such as distributed random forest, gradient Boosting, and deep learning. H2O also includes a Stacked Ensembles method that provides an optimal combination of a collection of prediction

algorithms using a process known as "stacking" (Candel et al. 2016)

Figure 2. X-Rite Ci60 portable spectrophotometer
Şekil 2. *X-Rite Ci60 taşınabilir spektrofotometre*

2.3. Naive Bayes

Naive Bayes is the main form of the Bayesian framework. Naive Bayes has efficient and effective learning algorithms which is one of the machine learning and data mining (Zhang 2004). Bayesian network classifier is a classification model based on a statistical method. In the Bayesian network classifier, the prior probability of the events is cleverly linked to the posterior probability (Fan et al. 2013). Naive-Bayes is an algorithm used to solve binary and multiclass classification problems, especially with excess data (Caruana and Niculescu-Mizil 2006; Lonita and Lonita 2018). It is generally used to determine the combined probability of words and classes, especially in the field of text mining (Amasyalı et al. 2006). It is a supervised, easy-to-use classification algorithm used for labeling and classifying data. Using the Bayesian theorem, it calculates the probability values of the effects of each criterion on the outcome and calculates which data belongs to which class (Çalış et al. 2013). Training and outcome procedures are very fast but fail to solve complex classification problems. Bayes' theorem is calculated by the following formula:

$$P(A/B) = (P(B/A) * P(A))/P(B) \ldots\ldots\ldots..(1)$$

In here;
P (A), the independent probability of event A (predominant probability),
P (B), independent probability of event B,
P (B | A) is the probability of event B (conditional probability)
P (A | B) is the probability of event A (conditional probability) (Çalış et al. 2013).

2.4. Random Forest

Bagging is a simple collection technique in which we build many independent determinants/models/learners and combine them using some average model techniques. We usually receive a random sub-sample / boot data for each model, so all models are slightly different from each other. Each observation is selected by modifying it to be used as input for each of the models. Thus, each model will have different observations based on the bootstrap process. This technique reduces errors by reducing variance, as it takes many unrelated students to make a final model. An example of a bagging community is a Random Forest model (Anonymous 2019). The Random Forest algorithm developed by Leo Bieman generates multiple trees to solve a question and creates different decision trees. Random Forest algorithm can be used in both classification and regression problems. Random Forest is an advanced version of the CART algorithm, in which many trees are created based on subsets of data. It is a supervised machine learning algorithm. Already, as the name suggests, it creates a forest and somehow makes it random. The forest is a collection of decision trees trained by the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall outcome. It is one of the most popular learning methods that provide simple, fast results in terms of understanding and implementation based on the aggregation of

estimates from multiple decision trees (Mitchell 2011). At each node, branches are formed according to the criteria of the CART algorithm (GINI index). The GINI index measures class homogeneity and can be expressed by the following formula (2) (Akar and Güngör 2012):

$$\sum\sum_{j\neq i}(f(C_i, T)/|T|)(f(Ci, T)/|T|) \ldots\ldots \quad .(2)$$

C_i and (f (C_i, T) shows the probability that the selected sample belongs to the ''Ci class'' (Akar and Güngör 2012).

2.5. Gradient Boosting Machines

Boosting is a community in which predictions are made sequentially, not independently. This technique uses the logic that subsequent estimators learn from the mistakes of previous estimators. Therefore, the likelihood of observations in subsequent models is uneven, and those with the highest errors appear most. The estimators can be selected from a range of models such as decision trees, regressors, classifiers, and so on. Since the new estimators learn from the mistakes of the previous estimators, it takes less time/iteration to get close to the actual predictions. But we have to choose the criteria carefully for stopping, or we can lead to an overload of training data. Gradient Boost is an example of the acceleration algorithm. Gradient Boosting Machines (GBM) is one of the powerful machine learning techniques. It shows significant success in a wide area of practical applications (Natekin and Knoll 2013). GBM is used for both regression and classification tree models. GBM regression and classification are forward-learning ensemble methods. It achieves predictive results by using gradually developed estimates. Boosting helps to improve the accuracy of trees and nonlinear regression (Click et al. 2017). GBM is a community where predictions are made in order, not independently. This technique uses the logic that subsequent estimators learn from the mistakes of previous estimators. Gradient Boost is an example of the acceleration algorithm (Anonymous 2019).

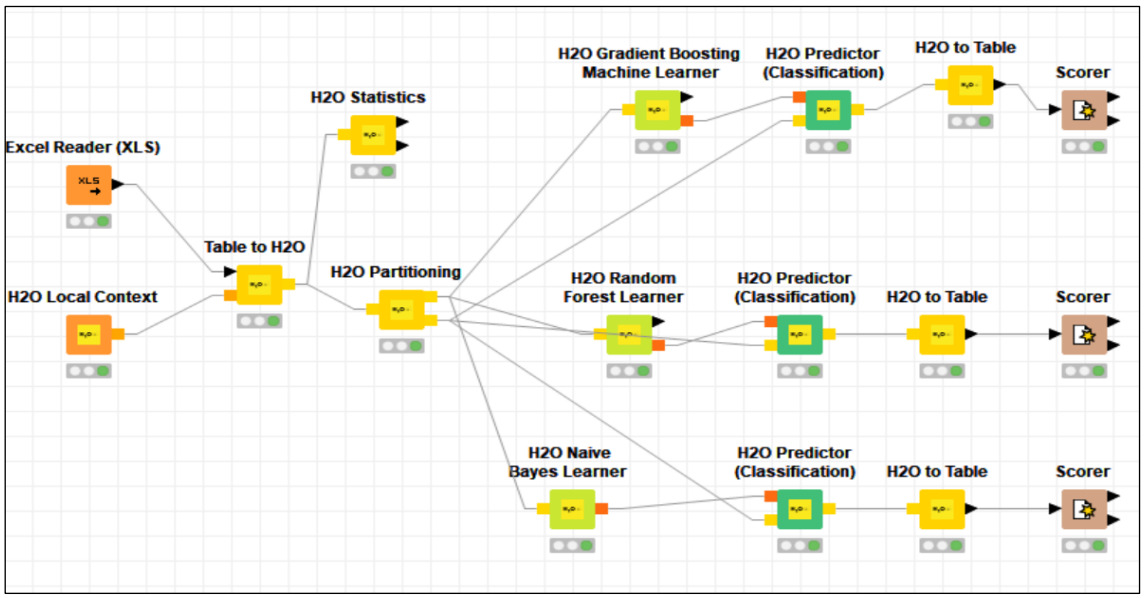

Figure 3. Workflow of KNIME
Şekil 3. *KNIME akış diyagramı*

2.6. Data Analysis

A database that contained 540 data of L*, a*, and b* values from the measurement of 180 fruits. The database was analyzed using descriptive statistical methods with the help of the KNIME Analytics Platform. For this purpose, H2O Naive Bayes, H2O Gradient Boosting Machine and H2O Random Forest node were

connected in the workflow of the KNIME Analytics Platform for the data analysis (Figure 3). The evaluation process of data, 70% of the data was used for training and 30% for the testing as a partitioning method. The following steps of the data analysis were expressed in Figure 4, respectively

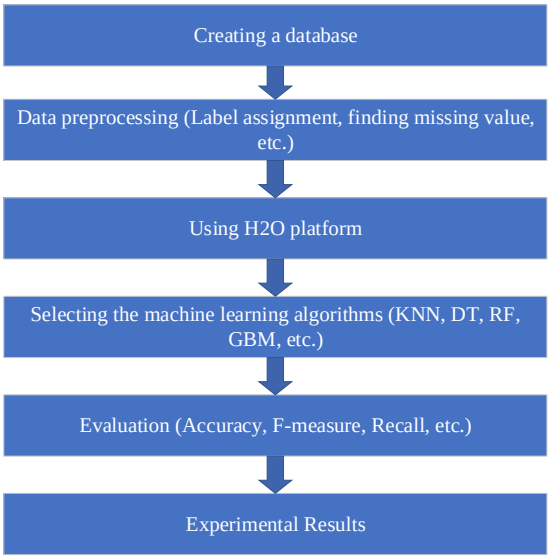

Figure 4. Steps of the data analysis
Şekil 4. *Veri analizi adımları*

3. Results and Discussion

The classifier performance, which accuracy, error, F-Measure, Cohen's Kappa, recall, precision, true positive (TP), false positive (FP), true negative (TN), false negative (FN) values were given in Table 2-3. As a result of the comparison, it was found that the Naive Bayes and Gradient Boosting Machine classification model was better than the Random Forest algorithm (Figure 4).

Table 2. The classifier performance of H2O machine learning analysis
Tablo 2. *H2O makine öğrenme analizinin sınıflandırıcı performansı*

Confusion Matrix	Algorithms	CC	WC	Accuracy (%)	Error (%)	FM	CK
Red Delicious		27	0	100	0	1	1
Golden Delicious	GBM	20	0	100	0	1	1
Granny Smith		17	0	100	0	1	1
Red Delicious		27	0	100	0	1	1
Golden Delicious	RF	20	0	100	0	0,976	1
Granny Smith		16	1	98,43	1,562	0,97	0,976
Red Delicious		27	0	100	0	1	1
Golden Delicious	Naive Bayes	20	0	100	0	1	1
Granny Smith		17	0	100	0	1	1

*CC: Correct classified, WC.: Wrong classified, FM: F-Measure, C K: Cohen's Kappa

In the literature, Kavdir and Guyer (2004) used fuzzy logic for apple grading. Classification results obtained by the fuzzy logic expert by 89%. Kleynen et al. (2005), using multi-light spectra, have made studies on multi-colored apples. Apples were examined in two groups, and 90% accuracy was achieved by using linear discriminator classifiers. Ronald and Evans (2016) used MATLAB software and the Naive Bayes algorithm for the classification of apple fruit varieties. The results showed that accuracy, sensitivity, precision, and specificity were 91%, 77%, 100%, and 80%, respectively. Sabancı et al. (2016) worked on classification parameters of apple varieties grown in the Karaman Region with the help of BayesNet, NaiveBayes, KStar, SMO, RBFNetwork, RBFClassifier, MLPClassifier, J48, RandomTree ve Random Forest algorithms. They found that according to size classification, the J48 algorithm had a 95.56% success rate, and the color classification MLPClassifier algorithm had a 97.78% success rate. Wu et al. (2017) studied the classification of apple varieties. They used near-infrared reflectance and fuzzy discriminant c-means clustering model (FDCM) for sorting apple

varieties. The clustering accuracy of FDCM achieved 97%. According to our results, techniques were used in this study 100 % success was obtained from two algorithms.

Table 3. The accuracy criteria of H2O machine learning analysis
Tablo 3. *H2O makine öğrenme analizinin doğruluk kriterleri*

Algorithms	Accuracy Criteria	Recall	Precision	Sensitivity	TP	FP	TN	FN
	Red Delicious	1	1	1	27	0	37	0
	Golden Delicious	1	1	1	20	0	44	0
H2O Naive Bayes	Granny Smith	1	1	1	17	0	47	0
	Red Delicious	1	1	1	27	0	37	0
	Golden Delicious	1	1	1	20	0	44	0
H2O GBM	Granny Smith	1	1	1	17	0	47	0
	Red Delicious	1	0,952	1	27	0	37	0
	Golden Delicious	1	1	1	20	1	43	0
H2O RF	Granny Smith	0,941	1	0,941	16	0	47	1

*TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative

4. Conclusion

Machine learning is very critical for the future of the agriculture industry. Because technological developments need to obtain databases from agricultural products for better machine automation systems. Also, it is not enough to get databases from agricultural products; additionally, the determination of optimum or best learning methods and mathematical models for this aim is essential. Because of this reason, machine learning techniques have been used in the classification of agricultural products from past to present. With the advancement of Artificial Intelligence every day, new techniques emerge. The future of the ML models goes to widespread of real-work applications. Nowadays, ML techniques are common with the repeated experiment but, still developing and scientists still learning the boundaries of ML. With the integration of machines, ML has a new future for automated decision-making or support to provide practical tools for knowledge-based agricultural applications. This development will result in better automation applications and machine control systems for agricultural production.

References

Aiello S, Eckstrand E, Fu A, Landry M & Aboyoun P (2016). Machine learning with R and H2O, http://h2o.ai/resources/ (Accessed to web: 31.08.2019).

Akar Ö & Güngör O (2012). Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. *Jeodezi ve Jeoinformasyon Dergisi*, s.139-146.

Amasyalı MF, Diri B, Türkoğlu F (2006). Farklı özellik vektörleri ile türkçe dokümanların yazarlarının belirlenmesi. The Fifteenth Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'2006), Muğla, Turkey, 21-24 June, 2006.

Anonymous (2019). Yapay Zeka, Robotik ve Sinirbilim. https://devhunteryz.wordpress.com/2018/07/11/gradyan-arttirmagradient-boosting/ (Accessed to web: 31.08.2019).

Bhatt AK, Pant D & Singh R (2014). An analysis of the performance of Artificial Neural Network technique for apple classification. AI & Society, 29(1): 103-111.

Bühlmann P (2012). Bagging, boosting and ensemble methods. In Handbook of Computational Statistics, Springer, pp. 985-1022, Berlin, Heidelberg.

Candel A, Parmar V, LeDell E & Arora A (2016). Deep learning with H2O. H2O. AI. Inc.

Canizo BV, Escudero LB, Pellerano RG, Rodolfo GW (2019). Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. *Computers and Electronics in Agriculture*, 162(2019):514–522.

Caruana R, Niculescu-Mizil A (2006). An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning, ACM, pp. 161-168.

Click C, Malohlava M, Parmar V, Roark H & Candel A (Nov 2017). Gradient Boosting Machine with H2O.

http://h2o.ai/resources/ (Accessed to web: 31.08.2019).

Çalış K, Gazdağı O, Yıldız O (2013). Reklam içerikli epostaların metin madenciliği yöntemleri ile otomatik tespiti. *Bilişim Teknolojileri Dergisi*, Cilt: 6, Sayı: 1, Ocak 2013.

Fan L, Huang X, Yi L (2013). Fault diagnosis for fuel cell based on naive bayesian classification. *TELKOMNIKA*, 11(12): 7664-7670, December 2013, e-ISSN: 2087-278X.

Kavdir I & Guyer DE (2004). Apple grading using fuzzy logic. *Turkish Journal of Agriculture and Forestry*, 27(6): 375-382.

Kleynen O, Leemans V, Destain M (2005). Development of a multi-spectral vision system for the detection of defects on apples. *Journal of Food Engineering*, 69: 41–49.

Lonita I, Lonita L (2018). Classification algorithms of data mining applied for demographic processes. *BRAIN – Broad Research in Artificial Intelligence and Neuroscience*, Volume 9, Issue1, February 2018, ISSN 2067-8957.

Mitchell MW (2011). Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters. *Open Journal of Statistics*, 2011(1): 205-211, doi:10.4236/ojs.2011.13024.

Mohsenin NN (1980). Physical properties of plant and animal materials. Gordon and Breach Science Publishers, One Park Avenue, New York 10016, p. 742, USA.

Nandi CS, Tudu B & Koley C (2014). Computer vision based mango fruit grading system. In International Conference on Innovative Engineering Technologies (ICIET 2014) Dec, pp. 28-29.

Natekin A & Knoll A (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7 (2013): 21.

Öztürk R (1988) Bazı meyve ve sebzelere uygun kombine tip boylama makinelerinin yapısal karakteristikleri, Doktora Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Tarımsal Mekanizasyon Anabilim Dalı, Ankara.

Ronald M & Evans M (2016). Classification of selected apple fruit varieties using Naive Bayes. *Indian Journal of Computer Science and Engineering*, 7(1): 13-19.

Sabancı K, Ünlerşen MF, Dilay Y (2016). Determination using image processing techniques the classification parameters of apple varieties grown in the Karaman region. *Journal of Agricultural Machinery Science*,12 (2), 133-139.

Seker SE & Erdogan D (2018). End-to-end data science with KNIME. 1. Press, p. 440, Demet Erdogan Publishing.

Semary NA, Tharwat A, Elhariri E & Hassanien AE (2014). Fruit-based tomato grading system using features fusion and support vector machine. *In Intelligent Systems'* 2014, pp. 401-410, Springer, Cham.

Sofu MM, Er O, Kayacan MC & Cetişli B (2016). Design of an automatic apple sorting system using machine vision. *Computers and Electronics in Agriculture*, 127: 395-405.

Suleiman D & Al-Naymat G (2017). SMS spam detection using H2O framework. Procedia computer science, 113: 154-161.

Wu X, Wu B, Sun J & Yang N (2017). Classification of apple varieties using near infrared reflectance spectroscopy and fuzzy discriminant c-means clustering model. *Journal of Food Process Engineering*, 40(2), e12355.

Zawbaa HM, Hazman M, Abbass M & Hassanien AE (2014). Automatic fruit classification using random forest algorithm. In 2014 14th International Conference on Hybrid Intelligent Systems, IEEE., pp. 164-168.

Zhang, H. (2004). The optimality of Naive Bayes. American Association for Artificial Intelligence, 1.2(2014): 3.