





## CLASSIFICATION OF RAISIN GRAINS VARIETY USING SOME MACHINE LEARNING METHODS

Yavuz UNAL <sup>1</sup>✉, Husamettin KAPLAN <sup>2</sup>, Yusuf BEKTAS <sup>2</sup>, Muhammed Bedirhan CAGLAR <sup>2</sup>

<sup>1</sup>Engineering Faculty, Amasya University, Amasya, Türkiye

<sup>2</sup>Technology and Innovation Department, Amasya University, Amasya, Türkiye

### Article History:

- received 10 November 2022
- accepted 03 May 2023

**Abstract.** One of the agricultural crops with considerable nutritional and financial worth is raisins. Every year, the world produces and consumes millions of tons of raisins. In this work, machine learning was used to categorize two different raisin kinds that are grown in our nation. Machine learning techniques Decision Trees and Random Forest were used to classify the 2-class data set with 7 different attributes that were acquired as a ready-made data set. With 020 Random Forest and Decision Trees, classification accuracy was 85.44% and 85.22%, respectively, in the analyses that were conducted.

**Keywords:** machine learning, random forest, decision trees, raisin grains, classification, artificial intelligence.

✉Corresponding author. E-mail: [yavuz.unal@amasya.edu.tr](mailto:yavuz.unal@amasya.edu.tr)

## Introduction

There are six nations with an annual grape production of 4 million tons or greater among those that practice viticulture. According to the most recent data, China produces 16% of the world's grapes, or around 74.5 million tons, followed by the USA, Italy, Spain, France, and Turkey. Spain makes up 13.1% of the 7.12 million hectares of vineyard land in the world, with China, France, Italy, Turkey, and the USA coming in second and third, respectively. Turkey has 467,093 acres of vineyard space, placing it fifth in the world, and produces 4,175,356 tons of fresh grapes, placing it sixth (Food and Agriculture Organization of the United Nations Statistics Division, 2022).

Turkey produces a significant amount of plants, and 1.83% of its land is used for viticulture cultivation. Figure 1 depicts the distribution of vineyard acreage and grape production levels in Turkey's seven geographical regions based on figures from 2016. Accordingly, the Aegean region accounts for around 52% of all grape production in our nation, followed by Southeastern Anatolia (18%), the Mediterranean (12%), Central Anatolia (8%), the Marmara (6%), Eastern Anatolia (3%), and the rest (1%). From the Black Sea region, it is provided (Soylemezoglu et al., 2015).

It can be challenging to distinguish between various food types and qualities in the food sector. This manual process is expensive and time-consuming. Additionally, this procedure won't be subjective. In recent years, image processing techniques have been successfully used to differentiate between the types and sizes of food (Francis & Clydesdale, 1975).

In this study, the Raisin Grain (Raisin Dataset, 2022) dataset, which is composed of features obtained from the images of two varieties of raisins grown in Turkey, has been classified with two of the machine learning algorithms.

Similar studies in this area are:

Cinar et al. (2020) used a camera in the system they created to take pictures of the raisins in the box. Three distinct machine methods were used to classify the obtained statistical data. The LR, MLP, and SVM algorithms were used to get classification results. Support vector machine (SVM), one of these techniques, had the greatest result with 86.44% (Cinar et al., 2020).

From photos of raisins, Okamura et al. (1993) extracted characteristics. To categorize these traits, they employed the naive Bayes method. Compared to the manual classification of raisins, they were more effective (Okamura et al., 1993).

Omid et al. (2010) created a mechanism that uses image processing techniques to extract the size and color features of raisins. Based on these extracted features, they performed classification and had a 96% success rate (Omid et al., 2010).

SVM was used by Yu et al. (2012) to categorize raisins. They categorize the raisins into four groups and achieve the greatest SVM classification success rate of 95% (Yu et al., 2012).

There are many known methods for the quality evaluation and classification of foods that are indispensable for human life. However, these traditional methods may not be efficient in terms of time and resources. In addition, the effect of the human factor in traditional methods can create negative effects such as inconsistency and inefficiency. These adverse events are a key factor in the development of more consistent methods to quickly and clearly assess the leading qualities of food products such as raisins. In addition, being able to distinguish the varieties in raisins determines in which area they can be used. Some grape varieties are used in the food industry as raw materials, while others can be consumed directly as snacks. For this reason, it is necessary to distinguish grape varieties from each other in terms of price and usage areas (Cinar et al., 2020). The aim of this study is to classify some of the raisin varieties grown in Turkey using machine learning methods. It is aimed to determine which machine learning algorithm performs better on this data set.

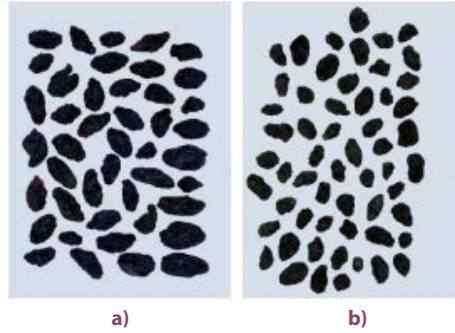
In the first part of the study, information about the characteristics of the data set, classification models used in the study and performance metrics are given. In the second part, the results obtained in the study are explained in detail. Discussion is given in the third section and discussion topics are given in the last section.

## 1. Material and methods

### 1.1. Dataset

The Raisin Dataset (Cinar et al., 2020). was employed in this study. There are 900 raisins' picture data with seven features in this data collection. This dataset has 2 classes. There are 450 Besni species and 450 Kecimen species grape data in it. The data set's sample photos are shown in Figure 1.

For each of the raisins present in the photos, several feature inferences were made during the feature extraction stage. The method of extracting features was done in terms of



**Figure 1.** Picture examples from which the data set was obtained: a – Besni; b – Kecimen

morphological traits. For every raisin grain, a total of 7 morphological traits were deduced. Many different image processing techniques use morphological feature inference to process images based on their forms. Each pixel in the image is altered throughout this process based on the values of the pixels surrounding it (Cinar et al., 2020).

There are 7 features in this dataset. These features are given in Table 1.

**Table 1.** Feature list of raisin grains

Feature	Min.	Mean	Max.	Std. Dev.
Area	25387	87804.128	235047	39002.111
Perimeter	619.074	1165.907	2697.753	273.764
MajorAxisLength	225.63	430.93	997.292	116.035
MinorAxisLength	143.711	254.488	492.275	49.989
Eccentricity	0.349	0.782	0.962	0.09
ConvexArea	26139	91186.09	278217	40769.29
Extent	0.38	0.7	0.835	0.053
Class	Kecimen	Besni		

MajorAxisLength gives the length of the main axis, which is the longest line of the grape grains.

MinorAxisLength gives the length of the main axis, which is the shortest line of the grape grains.

Eccentricity gives a measure of the eccentricity of the ellipse, which has the same moments as raisins.

ConvexArea gives the pixel count of the smallest convex skin of the region formed by the raisin.

Extent gives the ratio of the region boundaries formed by the raisin to the total pixels.

Area gives the number of pixels in a Raisin.

Perimeter measures the environment by calculating the distance between the boundaries of the raisin grain and the pixels around it.

The seven features in the data set and their histogram distributions are given in Figure 2.

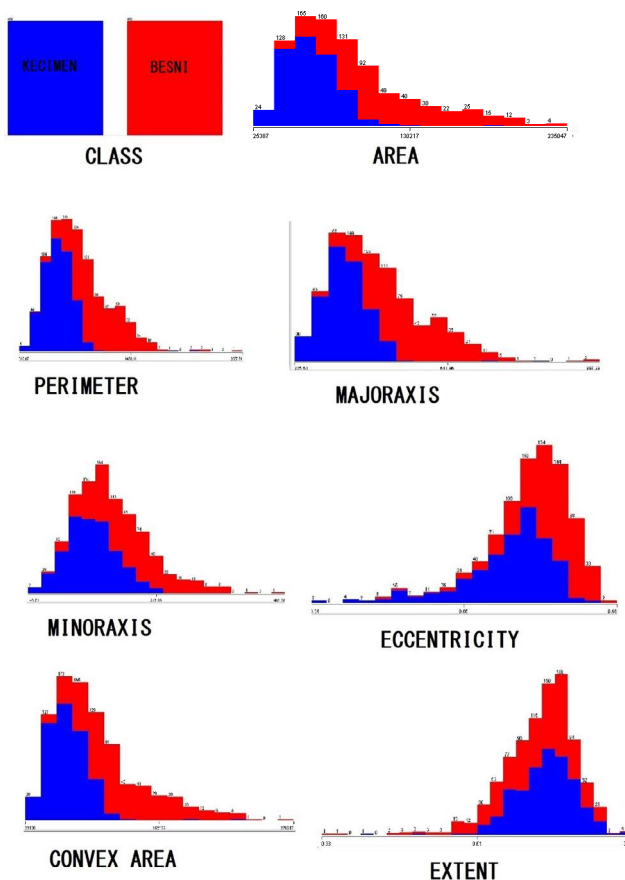


Figure 2. Histogram distribution of data set

## 1.2. Machine learning algorithms

Decision Trees and Random Forest are two machine learning algorithms employed in this study. Below are the theoretical justifications for these techniques. (Cinar et al., 2020). Logistic regression (LR), multi-layer perceptron (MLP) and Support vector machine (SVM) machine learning algorithms were applied on this data set. The reason we used Decision tree and Random forest algorithms in this study is that these algorithms have not been applied to this dataset before.

## 1.3. Decision trees algorithm

Because they are simpler to interpret than other classification methods, can be realized at lower costs, are simple to integrate with databases, and have a high level of reliability, decision trees are a common classification and regression methodology (Chein & Chen, 2008).

Additionally, this strategy works well with high-dimensional data and the leaves displayed as decision rules in decision trees are simple for experts in the field to interpret (Rokach & Maimon, 2008).

Learning and classification are the two stages of classification processing in decision trees. The so-called "training data set", in which the results matched to the values are known, is employed in the learning phase. The decision tree classification method receives the training data and analyses it to build the model. These models, which were discovered through analysis, are categorization rules or decision trees. The second stage, the classification step, begins as soon as the learning process is complete. The "test data set" data set is utilized in the classification phase. This process is used to evaluate the accuracy of any classification rules or decision trees that have been built (Chaudhuri et al., 1999).

#### 1.4. Random forest algorithm

Leo Breiman created the community machine-learning algorithm known as Random Forest. Instead of using a single classifier, ensemble classification techniques use many classifiers. The RF algorithm's structure consists of many decision trees, and it calculates outcomes by averaging the trees (Breiman, 2001).

Two parameters form the basis of the RF. These settings determine how many trees will be produced (N) and how many random variables will be used to fill each node (m). When building regression trees or classification trees, the assumed m value was suggested as  $p/3$  and indicates the total number of predictive variables.

By lowering the correlation between trees without significantly raising the variance, the random forest approach aims to enhance the variance reduction of the bagging method. This is accomplished by selecting input variables for tree growth at random (Cutler et al., 2007).

#### 1.5. Performance metrics

There are certain standards for measuring the effectiveness of the machine learning techniques employed on the data set for categorization. The confusion matrix, which is shown as a  $2 \times 2$  matrix, represents the success of the model's prediction accuracy. It offers a comparison between the actual values and the forecasts (Schaffer, 1993).

Accuracy: The percentage of values that were successfully predicted to all other values. The equation is shown below (Cinar et al., 2020):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

Precision: It is the ratio of correctly predicted values to the sum of incorrectly predicted values and correctly predicted values. The formula is given below (Cinar et al., 2020):

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Sensitivity: It is the ratio of correctly predicted values to the sum of correctly predicted values and incorrectly predicted values. The formula is given below (Cinar et al., 2020):

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

Specificity: It is the ratio of True Negative to the sum of true negative and false positive. The formula is given below (Cinar et al., 2020) (4).

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

F-Measure: It is the value given to us by the precision and sensitivity values, whose harmonic average is taken. It is a measure of the precision and robustness of the model. The formula is given below (Cinar et al., 2020):

$$F - Measure = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

## 2. Results

In this study, performance measurements were obtained on the raisin dataset using 2 different machine-learning algorithms.

Cross-validation is an error prediction technique created with the goal of enhancing classification security. The dataset is randomly divided into a predetermined number of subsets for training and testing during cross-validation. One of the subsets is accepted as a test set, and the system is trained using the other sets. The system is tested after this process is done for as many data sets as necessary (Gupta, 2017).

In this study, test and training datasets were created using 10 cross-fold validation.

The confusion matrices obtained as a result of the classification are given in Table 2 and Table 3, and the statistical data obtained from here is used as a performance measure.

**Table 2.** The confusion matrix of the Random Forest algorithm

Random Forest Algoritihm	Predicted Kecimen	Predicted Besni
Actual Kecimen	399	51
Actual Besni	80	370

**Table 3.** The confusion matrix of the Decision Tree algorithm

Decision Tree	Predicted Kecimen	Predicted Besni
Actual Kecimen	389	61
Actual Besni	72	378

The approach that provides the best accuracy-based value, according to the average classification based on Table 3, which shows performance measures and comparisons for each method and class, is Random Forest, with a value of 85.44%. The Decision Tree produced a classification accuracy of 85.22% when used on the same dataset. In Table 4, data for sensitivity, specificity, precision, and F-Measure are also displayed.

**Table 4.** Classification and performance results.

Performance Measure	Random Forest	Decision Tree
Accuracy	85.44%	85.22%
Sensitivity	83.30	84.38
Specificity	87.89	8.10
Precision	85.60	85.20
F-Measure	85.40	85.20

### 3. Discussion

Higher classification achievements can be reached by increasing the number of image sets and adding morphological features along with color layer, form, and texture data in addition to the features gathered from the products, according to an analysis and evaluation of the study's results.

Different machine learning approaches or hybrid models can be developed and studied on the pertinent data set in addition to the machine learning techniques used in the current study.

In this study, unlike the previous ones, two previously unused machine learning algorithms were applied to this data set. These are decision tree and random forest algorithms. Among the two classification algorithms applied in this study, the Random forest algorithm gave the better result. Classification using the random forest algorithm has achieved 85.44% classification success. It outperformed the classification success of 84.22% with logistic regression in the previous study.

### Conclusions

In this study, Kecimen and Besni raisin varieties produced and exported in Turkey were classified using random forest and decision tree, which are machine learning algorithms. According to the classification results, the random forest algorithm performed better. It has been observed that machine learning algorithms have been successfully applied to such agricultural data sets. In future studies, various machine learning algorithms can be applied on other raisin varieties or other agricultural datasets.

### References

- Breiman, L. (2001) Random forest. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chaudhuri, S., Fayyad, U., & Bernhardt, J. (1999). Scalable classification over SQL databases. In *Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337)* (pp. 470–479). IEEE. <https://doi.org/10.1109/ICDE.1999.754963>
- Chein, C. F., & Chen, L. F. (2008) Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications*, 34, 280–290. <https://doi.org/10.1016/j.eswa.2006.09.003>
- Cinar, I., Koklu, M., & Tasdemir, S. (2020). Classification of raisin grains using machine vision and artificial intelligence methods. *Gazi Journal of Engineering Sciences*, 6(3), 200–209. <https://doi.org/10.30855/gmbd.2020.03.03>

- Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783–2792. <https://doi.org/10.1890/07-0539.1>
- Food and Agriculture Organization of the United Nations Statistics Division. (2022, July 14). *Crops and livestock products*. <http://www.fao.org/faostat/en/#data/qc>
- Francis, F. J., & Clydesdale, F. M. (1975). *Food colorimetry: Theory and applications*. AVI Publishing, Westport.
- Gupta, P. (2017, June 5). *Cross-validation in machine learning*. <https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f>
- Okamura, N. K., Delwiche, M. J., & Thompson, J. F. (1993). Raisin grading by machine vision. *Transactions of the ASAE*, 36(2), 485–492 <https://doi.org/10.13031/2013.28363>
- Omid, M., Abbasgolipour, M., Keyhani, A., & Mohtasebi, S. S. (2010). Implementation of an efficient image processing algorithm for grading raisins. *International Journal of Signal Image Processing*, 1(1), 31–34.
- Raisin Dataset. (2022). [Data set]. <https://www.muratkoklu.com/datasets/>
- Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: Theory and applications*. World Scientific. <https://doi.org/10.1142/6604>
- Schaffer, C. (1993). Selecting a classification method by cross-validation. *Machine Learning*, 13(1), 135–143. <https://doi.org/10.1007/BF00993106>
- Soylomezoglu, G., Kunter, B., Akkurt, M., Sağlam, M., Ünal, A., Buzrul, S., & Tahmaz, H. (2015). Viticulture development methods and production targets. In *Turkish Agricultural Engineering 8th Technical Congress, Proceedings* (pp. 606–629).
- Yu, X., Liu, K., Wu, D., & He, Y., (2012). Raisin quality classification using least squares support vector machine (LSSVM) based on combined color and texture features. *Food Bioprocess Technology*, 5(5), 1552–1563. <https://doi.org/10.1007/s11947-011-0531-9>