All Sciences Proceedings
http://as-proceeding.com/

5th International Conference on Applied Engineering and Natural Sciences

July 10-12, 2023 : Konya, Turkey

https://www.icaens.com/

Classification of Poisonous and Edible Mushrooms with Optimized Classification Algorithms

Sedat Metlek*[1], Halit Çetiner [2]

[1]Vocational School of Technical Sciences, Burdur Mehmet Akif Ersoy University, Türkiye
[2]Vocational School of Technical Sciences, Isparta University of Applied Sciences, Türkiye

*(sedatmtelek@mehmetakif.edu.tr)

Abstract – Within the scope of this study, it is aimed to classify the mushroom species consumed as a staple food. For this purpose, 8124 mushroom data with 22 different mushroom feature information were used. 5686 of these data were used for training and 2438 for testing. In the study, poisonous and edible mushroom species were classified by random forest, decision tree, and logistic regression classification methods. The parameters used in the random forest and decision tree classification algorithms used in the study were optimized with the GridSearchCV optimization method. With the random forest algorithm, the highest precision, recall, and F1 score values are 0.93, 0.98, and 0.95, respectively. When these values are examined on a class basis, the highest distinctiveness results were obtained in the poisonous class. In the edible class, the highest performance results were measured as 0.97, 0.92, and 0.95 for precision, recall, and F1 score values, respectively. With the decision Tree algorithm, the highest precision, recall, and F1 score values are 0.98, 0.98, and 0.92, respectively. The highest precision, recall, and F1 score values of the best poisonous class are 0.90, 0.98, and 0.92, respectively. The best performance results of the edible class were obtained with the highest precision, recall, and F1 score values of 0.98, 0.89, and 0.90, respectively. The average accuracy rate was 0.9028 with the Logistic Regression algorithm, and the precision, recall, and F1 score values of the poisonous class were obtained as 0.86, 0.97, and 0.91, respectively. Precision, recall, and F1 score values of the Edible class were obtained as 0.96, 0.83, and 0.89, respectively.

Keywords – Random Forest, Decision Tree, GridSearchCV, Logistic Regression, Mushrooms

I. INTRODUCTION

The discovery and consumption of mushrooms as a food type by human beings dates back to the first ages [1]. This highly satisfying food source contains amino acids, carbohydrates, fiber, important vitamins, and minerals. Mushrooms are also a frequently used resource in the pharmaceutical industry [2]. Fungi species are divided into 3 groups according to their nutrition types: Mycorrhizal (Symbiotic), Saprotrophic (Saprophytes), and Parasites. Mycorrhizal species usually live mutualistically with a host plant. The saprotrophic species produce their food from dead organic materials. Some varieties of this species form the basis of cultivated mushrooms. Parasite fungi, on the other hand, provide their food by establishing a non-symbiotic relationship with other living things. There are about 145 groups of fungi as parasite species [3]. Apart from these, mushrooms are generally divided into two groups medicinal (poisonous) and edible mushrooms. Edible mushrooms are mushrooms that can be consumed as fresh or dried fruit parts.

These mushrooms have nourishing, stress-relieving, and anti-infective properties. Poisonous mushrooms are generally used in pharmaceutical applications due to the bioactive components and triterpenoids they contain. It is also used in the cosmetic industry. Poisonous mushrooms are not consumed directly [4].

A significant part of the mushrooms consumed in the world is still supplied directly from nature. Especially in rainy seasons, there is a significant increase in the number of mushrooms found in nature. However, most of the mushrooms obtained from nature are poisonous mushrooms. However, this type of mushroom is often confused with edible mushrooms. Because, whether the mushrooms are medicinal, i.e. poisonous or edible, is based primarily on visual identification and then on biochemical analyzes [5].

It is a very difficult and dangerous issue for people who are not experts in this field to come to a conclusion on this subject by doing biochemical analysis in daily life. In such cases, making the wrong decision often results in death or disability.

The main focus of the study is the development of software that can be used in the laboratory environment to distinguish between edible mushrooms and poisonous mushrooms. In this sense, the literature contribution of the study;

- Edible and poisonous mushrooms are classified using random forest(RF), decision tree(DT), and logistic regression(LR) classification algorithms.
- The success results of each classification algorithm used in the study were compared with each other.
- The parameters used with the decision tree algorithm were optimized using the GridSearch optimization algorithm, thus improving the performance results.

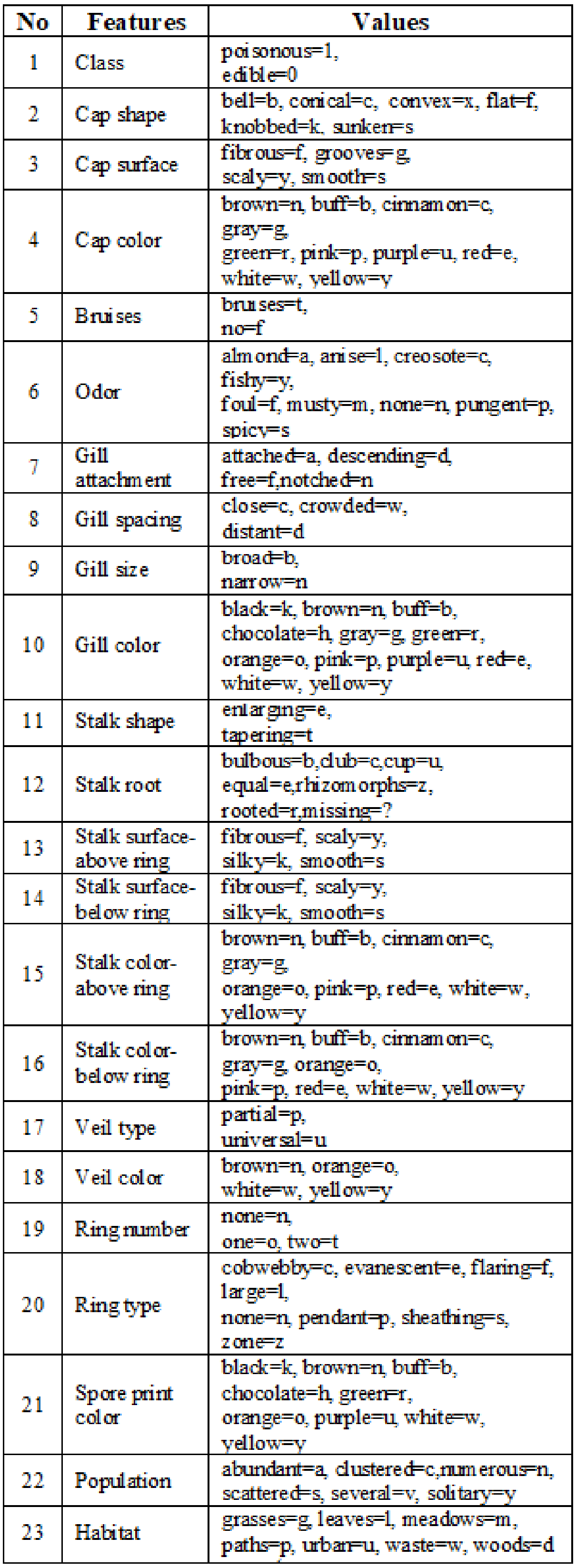
In the next section of the study; In section 2, information about the dataset used in the study is presented. In section 3, information about RF, DT, and LR in the literature is shared and information about optimized DT is presented. In addition, the evaluation metrics used in the study are also shared in section 3. In the last section of the study, the experimental success of each classification algorithm was shared separately and a general evaluation of the study was made.

II. Materials

In the study, the mushroom dataset prepared by the "Audobon Society Field Guide", which has been used for a long time in the literature, was used to classify edible and poisonous mushrooms [6].

The most important purpose of choosing this dataset is to enable users to test the optimization process in the DT algorithm in the study. There are 22 attributes belonging to edible and poisonous mushroom classes in the dataset used.

Table 1. Example of a table



No	Features	Values
1	Class	poisonous=1, edible=0
2	Cap shape	bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
3	Cap surface	fibrous=f, grooves=g, scaly=y, smooth=s
4	Cap color	brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
5	Bruises	bruises=t, no=f
6	Odor	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
7	Gill attachment	attached=a, descending=d, free=f,notched=n
8	Gill spacing	close=c, crowded=w, distant=d
9	Gill size	broad=b, narrow=n
10	Gill color	black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
11	Stalk shape	enlarging=e, tapering=t
12	Stalk root	bulbous=b,club=c,cup=u, equal=e,rhizomorphs=z, rooted=r,missing=?
13	Stalk surface-above ring	fibrous=f, scaly=y, silky=k, smooth=s
14	Stalk surface-below ring	fibrous=f, scaly=y, silky=k, smooth=s
15	Stalk color-above ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
16	Stalk color-below ring	brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
17	Veil type	partial=p, universal=u
18	Veil color	brown=n, orange=o, white=w, yellow=y
19	Ring number	none=n, one=o, two=t
20	Ring type	cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
21	Spore print color	black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
22	Population	abundant=a, clustered=c,numerous=n, scattered=s, several=v, solitary=y
23	Habitat	grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

These attributes are basic attributes that can be easily identified by anyone encountering a fungus in nature. The attributes of the dataset used are presented in detail in Table 1. In the dataset used, the attributes of 8124 mushrooms belonging to the poisonous, and edible mushroom species are shown in Table 1. This data was divided into two parts: 70% training and 30% testing. As a result of this process, 5686 data were used as training data and 2438 as test data.

III. Materials

In this part of the study, general information about RF, DT, and LR algorithms used for classification is presented. Afterward, the optimized DT and RF algorithms are mentioned.

A. *Decision Tree*

Decision trees are a popular classification algorithm used in many different applications. It has many different models such as ID3, and C4.5. The model used in the study is the C4.5 model developed by J. Ross Quinlan [7]. The random forest algorithm, which is extremely easy to understand, consists of two important steps: tree creation and pruning.

The general structure of a decision tree consists of leaves, branches, and roots. The bottom part of the tree structure is called the leaf and the top part is called the root. Each feature in the dataset represents node points. The structure or branch between two nodes is called. Deciding on which attribute value to branch is the most important step in constructing decision trees [8]. Generally, the gini index, information gain, and towing rule [3] are used as decision-making conditions [7], [9]. Gini information gain was used as a decision-making condition from the decision tree model used in the study. In this method, the effect of the related attribute on the result is calculated with an entropy-based value for each features.

If it is thought that there are n classes in a decision tree structure and it is thought to repeat these classes T times, the probability of data belonging to this class is calculated by Equation 1.

$$P_i = \frac{c_i}{|T|} \tag{1}$$

The c_i value in Equation 1 represents the class value of a class. The entropy $H(T)$ value of these classes is calculated by Equation 2.

$$H(T) = \sum_{i=1}^{n} P_i log_2 \tag{2}$$

According to the Y attribute value in the dataset, when the T class values are subclassed as $T_1, T_2, \dots\dots. T_n$ the information gain according to the Y attribute value is calculated using $IG(Y, T)$ Equation 3.

$$IG(Y,T) = H(T) - \sum_{i=1}^{n} \frac{|T_1|}{|T|} H(T_i) \tag{3}$$

In determining the feature value, the separation information is calculated using Equation 4.

$$SI(Y) = -\sum_{i=1}^{n} \frac{|T_1|}{|T|} log_2 \left(\frac{|T_i|}{T}\right) \tag{4}$$

The ratio of the information gain to the separation information gives the information about how much information gain will be provided by the separation of the relevant attribute. Similarly, the tree structure is separated according to the feature with the highest gain information by calculating the gain information for each feature.

Another important process used in the structure of decision trees is pruning. Pruning can be done in two ways [10]. When the tree structure is formed, when the tree grows at a certain rate, it is called pre-pruning to stop the division so that the tree does not grow anymore. Secondly, pruning by calculating the split points created after the tree is fully formed is called final pruning.

B. *Random Forest*

The random forest (RF) algorithm was first converted into a classification and regression tool by Breiman to make predictions based on various variables [11]. This algorithm is a classification algorithm that includes several decision trees constructed based on randomly selected subsets using bootstrap aggregating (bagging) [12]. It randomly selects the samples used during the training of the algorithm. At the same time, it randomly selects the nodes of the decision tree when dividing [13]. Besides randomness, it controls parameters such as forest size, structure, and node size, which are the parameters used in the construction of trees. Among these parameters, especially entropy, gini, and depth number parameters were used in different combinations in training the model.

If RF aims to predict future data, classification, and regression rules need to be established correctly [14]. The main aim here is to provide optimization that will provide high performance by minimizing the classification error rate. With this parameter optimization, the relationship between the used features and the class was investigated correctly and a good result was tried to be obtained.

The main steps that build the RF algorithm consist of the following steps respectively [14]:

- Based on the M value, randomly identify new sub-attribute sets named θ_k. θ_k is independent of any other subsets in the $\theta_1, \ldots, \theta_k$ sequence.
- Make individual decisions by training each of the subsets. Each subclassifier is represented by $h(X, \theta_k)$. The X value here represents the entries.
- The RF classifier is defined by repeating the above processes until values are obtained from all feature subsets.
- Finally, the class label is decided according to the results obtained from each classification.

C. *Logistic Regression*

Logistic regression is a non-linear form of a linear regression model. It usually calculates the probability of class membership for one of the two categories in the dataset and is useful when the dependent variable is restricted to a two-class problem [13]. In the study, there are two classes poisonous and edible. This aspect is a very suitable classification model for the application determined by logistic regression. The relationship between the variables using logistics can be expressed by Equation 5 [15], [16].

$$\bar{p} = \frac{1}{1 + e^{-\bar{\mu}}}, where\ \bar{\mu} = \bar{\theta}.x \qquad (5)$$

$\bar{\mu}$ is a linear function of x. $\bar{p}$ is used to estimate poisonous and edible classes.

D. *GridsearchCV*

In general, models developed in machine learning applications are trained on a dataset, and then the best-performing one is selected. However, we cannot say for sure which of these models is the best when different situations are involved. Therefore, the general aim is the continuous improvement of the models. In addition, a factor affecting the performance of the models is the selected parameters. For this reason, the use of a developed or used model with optimum parameter values directly affects the success of the system.

It is one of the methods used for this purpose in GridSearchCV. Grid Search is a well-known method for identifying all combinations of hyperparameters. The learning rate and the number of layers are the two most important parameters in GridSearch. First, a set of values is determined for each hyperparameter. The hyperparameter combination is determined in each loop. In the end, the most successful combination of the hyperparameters is selected and used in the learning process [17]. The optimized parameters determined for random forest and decision tree in the study are presented in detail in Tables 2 and 3.

Table 2. Selection results of random forest algorithm training parameters with GridSearchCV method.

Model	Criterion	Max-depth	Max-features	N estimators
Random Forest	Gini, Entropy	4, 5, 6, 7, 8	'auto', 'sqrt', 'log2'	200, 500
GridSearch CV	Gini	8	'auto'	500

Table 3. Selection results of decision tree algorithm training parameters with GridSearchCV method

Model	Criterion	Max-depth	Splitter	Max-features
Decision Tree	Gini, Entropy	1, 2, 3, 4, 5	‘best’ ‘random’	'auto', 'sqrt', 'log2'
GridSearch CV	Gini	5	'best'	'sqrt'

E. *Evaluation Metrics*

Accuracy, Recall, Precision, and F1 Score performance evaluation metrics, which are preferred in many applications in the literature, were used to evaluate the success of the models used in the study. These metrics used are presented in detail in Equations 6-9. In addition to these, a complexity matrix was also created for the outcome of each model [18]–[22].

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$F1\ Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{9}$$

IV. RESULTS AND CONCLUSION

In this study, the classification of fungi was carried out using 22 characteristic features and three different classification algorithms found in the literature. Evaluation metrics, which are also widely used in the literature, were used to measure classification success. The test performance results obtained for random forest, decision tree, and logistic regression are presented in detail in Tables 4, 5, and 6, respectively. In addition to these, the confusion matrix of each model is shared below.

Table 4. Selection results of the Decision tree algorithm training parameters with the GridSearchCV method. P=Poisonous, E=Edible

Model	Criterion	Class	Precision	Recall	F1 Score
Random Forest (Optimization)	Gini	P	0.93	0.98	0.95
Random Forest (Optimization)	Gini	E	0.97	0.92	0.95
Average Accuracy	0.9400				
Random Forest	Entrophy	P	0.91	0.96	0.93
Random Forest	Entrophy	E	0.96	0.89	0.93
Average Accuracy	0.9300				
Random Forest	Gini	P	0.91	0.96	0.93
Random Forest	Gini	E	0.96	0.90	0.92
Average Accuracy	0.9295				

The average accuracy obtained from the optimized parameters shown in Fig. 3 outperformed the results with entropy criteria in Fig. 1 and 1.05% better than the results with gini criteria in Fig. 2.

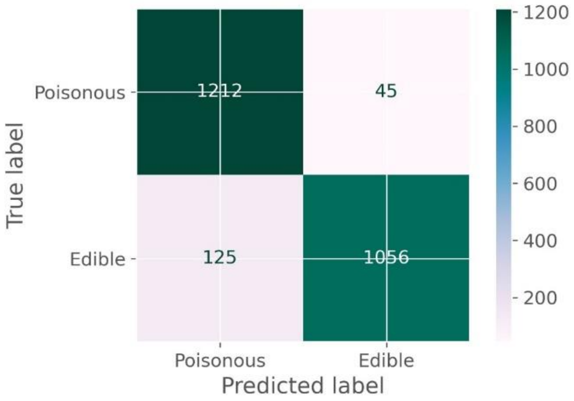


Fig. 1 Performance results of random forest algorithm with entropy type

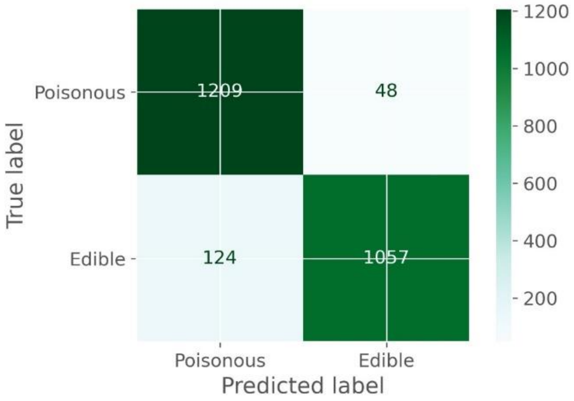


Fig. 2 Performance results of Random Forest algorithm with gini type

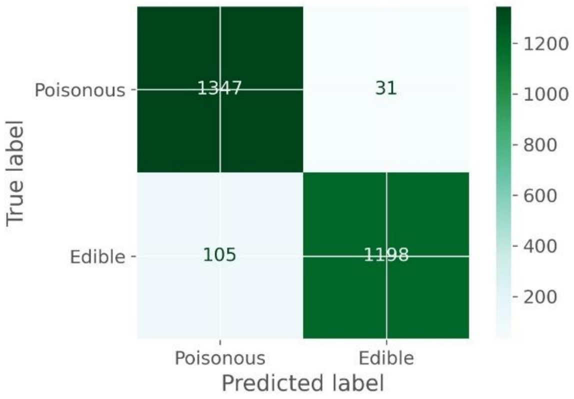


Fig. 3 Optimized performance results of random forest algorithm

As can be seen from the confusion matrix results in Figs 1, 2, and 3 in the random forest algorithm, a noticeable increase in classification success has been achieved with optimized parameter values.

Table 5. Decision Tree classification results. P=Poisonous, E=Edible

Model	Criterion	Class	Precision	Recall	F1 Score
Decision Tree (Optimization)	Entrophy	P	0.87	0.98	0.92
Decision Tree (Optimization)	Entrophy	E	0.98	0.84	0.90
Average Accuracy	0.9126				
Decision Tree	Entrophy	P	0.90	0.91	0.90
Decision Tree	Entrophy	E	0.90	0.89	0.90
Average Accuracy	0.8995				
Decision Tree	Gini	P	0.89	0.90	0.90
Decision Tree	Gini	E	0.89	0.90	0.90
Average Accuracy	0.8946				

In Figs 4, 5, and 6, the performance results of the Decision Tree algorithm obtained with different criteria are shown. Figure 4 shows the test performance results from the model trained using the optimized parameters.

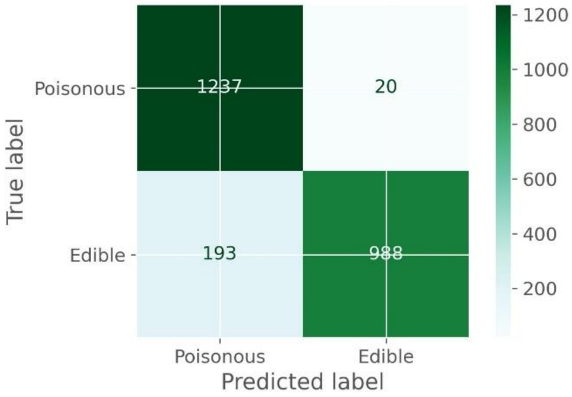


Fig. 4 Performance results of the decision tree algorithm with optimized parameters

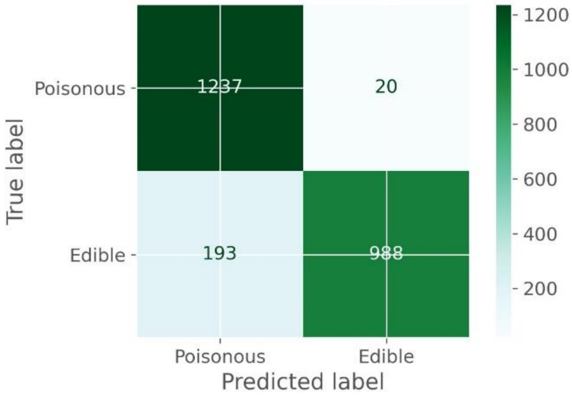


Fig. 5 Performance results obtained with the entropy type type of the Decision tree algorithm.

In Fig 5, the test performance result of the criterion-trained model with the entropy criterion feature is given. Fig 6 shows the performance result of the model trained with the gini criterion parameter.

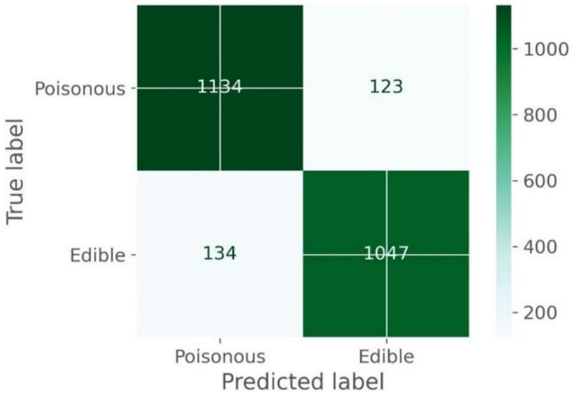


Fig. 6 Performance results obtained with the gini type of the decision tree algorithm

When Figs 4.5 and 6 are examined, it is seen that the performance results are quite close to each other.

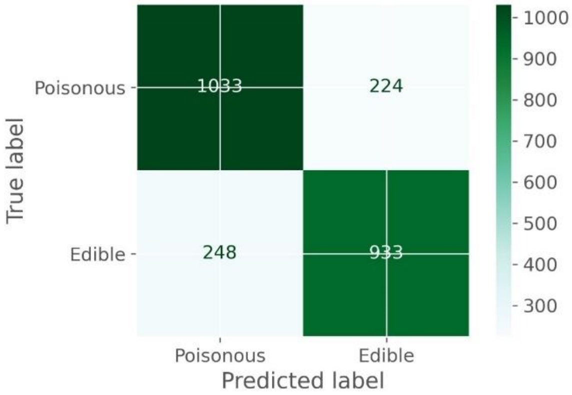


Fig. 7 Performance results of the Decision Tree algorithm with optimized parameters.

Finally, the performance results obtained using the logistic regression algorithm are given in Table 6 and Fig. 8. In terms of average accuracy, it can compete with the decision tree algorithm. However, a lower result was obtained than the performance results obtained from the random forest algorithm.

Table 6. Logistic Regression classification results. P=Poisonous, E=Edible

Model	**Class**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	P	0.87	0.98	0.92
Logistic Regression	E	0.98	0.84	0.90
Average Accuracy	0.9126			

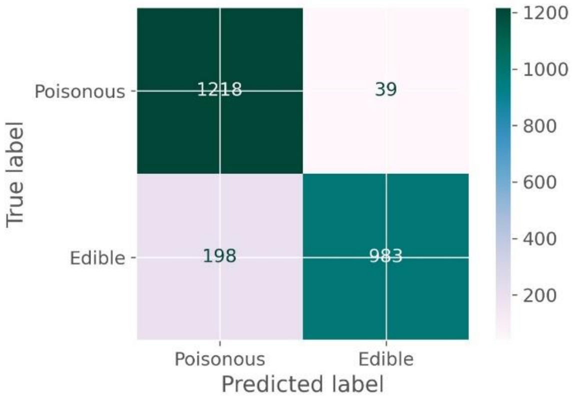


Fig. 8 Logistic Regression algorithm performance results.

With the Random Forest algorithm, the highest precision, recall, and F1 score values are 0.93, 0.98, and 0.95, respectively. When these values are examined on a class basis, the highest distinctiveness results were obtained in the poisonous class. In the Edible class, the highest performance results are 0.97, 0.92, and 0.95 for high precision, recall, and F1 score values, respectively.

With the decision tree algorithm, the highest precision, recall, and F1 score values are 0.98, 0.98, and 0.92, respectively. The highest precision, recall, and F1 score values of the poisonous class are 0.90, 0.98, and 0.92, respectively. The best performance results of the Edible class were obtained with the highest precision, recall, and F1 score values of 0.98, 0.89, and 0.90, respectively.

In order to obtain better test performance results given above, it is necessary to determine the best of the features. Better performance results can be obtained by removing the dataset from unnecessary attributes.

REFERENCES

[1] B. A. Wani, R. H. Bodha, and A. H. Wani, "Nutritional and medicinal importance of mushrooms," *J. Med. plants Res.*, vol. 4, no. 24, pp. 2598–2604, 2010.

[2] P. Kalač, "A review of chemical composition and nutritional value of wild-growing and cultivated mushrooms," *J. Sci. Food Agric.*, vol. 93, no. 2, pp. 209–218, Jan. 2013, doi: https://doi.org/10.1002/jsfa.5960.

[3] G. Y. Turp and M. Boylu, "Tıbbi ve Yenilebilir Mantarlar & Et Ürünlerinde Kullanımı," *Yuz. Yıl Univ. J. Agric. Sci.*, vol. 28, no. 1, pp. 144–153, 2018.

[4] S. M. Badalyan, N. G. Gharibyan, and A. E. Kocharyan, "Erspectıve in Usage of Bıoactıve Substances of Medıcınal Mushrooms in Pharmaceutıcal and Cosmetıc Industy," 2007.

[5] K. Tutuncu, I. Cinar, R. Kursun, and M. Koklu, "Edible and poisonous mushrooms classification by machine learning algorithms," in *2022 11th Mediterranean Conference on Embedded Computing (MECO)*, 2022, pp. 1–4.

[6] U. M. L. Repository, "Mushrooms." 1987, doi: https://doi.org/10.24432/C5959T.

[7] J. R. Quinlan, "Program for machine learning," *C4. 5*, 1993.

[8] T. Kavzoğlu and İ. Çölkesen, "Karar ağaçları ile uydu görüntülerinin sınıflandırılması," *Harit. Teknol. Elektron. Derg.*, vol. 2, no. 1, pp. 36–45, 2010.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees Belmont," *CA Wadsworth Int. Gr.*, 1984.

[10] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man. Mach. Stud.*, vol. 27, no. 3, pp. 221–234, 1987.

[11] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[12] W. Wang, X. Zhang, S.-H. Wang, and Y.-D. Zhang, "Covid-19 diagnosis by WE-SAJ," *Syst. Sci. Control Eng.*, vol. 10, no. 1, pp. 325–335, Dec. 2022, doi: 10.1080/21642583.2022.2045645.

[13] Y. Sun, H. Cheng, S. Zhang, M. K. Mohan, G. Ye, and G. De Schutter, "Prediction & optimization of alkali-activated concrete based on the random forest machine learning algorithm," *Constr. Build. Mater.*, vol. 385, p. 131519, 2023, doi: https://doi.org/10.1016/j.conbuildmat.2023.131519.

[14] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 3,

p. e1301, May 2019, doi: https://doi.org/10.1002/widm.1301.

[15] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review,” *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, 2002.

[16] J. Berkson, “Application of the Logistic Function to Bio-Assay,” *J. Am. Stat. Assoc.*, vol. 39, no. 227, pp. 357–365, Sep. 1944, doi: 10.1080/01621459.1944.10500699.

[17] F. M. J. M. Shamrat, P. Ghosh, M. H. Sadek, M. A. Kazi, and S. Shultana, “Implementation of machine learning algorithms to detect the prognosis rate of kidney disease,” in *2020 IEEE international conference for innovation in technology (INOCON)*, 2020, pp. 1–7.

[18] P. Ghosh *et al.*, “Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques,” *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

[19] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, “MIFH: A machine intelligence framework for heart disease diagnosis,” *IEEE access*, vol. 8, pp. 14659–14674, 2019.

[20] L. Wang, W. Zhou, Q. Chang, J. Chen, and X. Zhou, “Deep ensemble detection of congestive heart failure using short-term RR intervals,” *IEEE Access*, vol. 7, pp. 69559–69574, 2019.

[21] F. Miao, Y.-P. Cai, Y.-X. Zhang, X.-M. Fan, and Y. Li, “Predictive modeling of hospital mortality for patients with heart failure by using an improved random survival forest,” *IEEE Access*, vol. 6, pp. 7244–7253, 2018.

[22] S. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” *IEEE access*, vol. 7, pp. 81542–81554, 2019.