International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 4, Issue 4, April – 2020, Pages: 1-7

Determination of the Properties That are Effective in Determining the Edible State of Mushroom (Agaricus and Lepiota) by Using Datamining Methods

Oznur Isci Guneri[1], Burcu Durmus[2]*
Department of Statistics, Faculty of Statistics, Mugla Sitki Kocman University, Mugla, Turkey
Email: oznur.isci@mu.edu.tr[1], burcudurmus@mu.edu.tr[2]

Abstract: *The aim of this study is to determine the mushroom characteristics that are effective in classification with the help of decision trees by identifying data mining algorithms that can classify mushrooms according to their edibility status. In the scope of classification analysis, 22 different characteristics of mushrooms were examined. Classification analysis was performed with different decision tree algorithms and decision tree structures were examined. All of the algorithms discussed were very successful according to correct classification rate, Kappa statistics and root mean square error statistics. On the other hand, when the decision tree structures of the algorithms were examined, it was seen that the classification structures highlight different fungal properties and various comparisons were made in this direction. As a result of the study; the decision tree algorithms were found to be successful in detecting edibility in two types of mushroom families and the 'odor' attribute was found to be more distinctive than the other characteristics in the determination of edibility. In addition, it was found that 7 out of 22 attributes did not play any role in distinguishing edible mushroom.*

Keywords— classification; datamining; decision tree; mushroom

1. INTRODUCTION

Mushrooms are one of the plants with the most diversity in the world. It is an important plant flora used both as a nutrient and in alternative treatment. It is estimated that there are 1.5 million fungal species in the world [1]. Mushrooms, which can easily spread in many natural areas, are very popular as food with their delicious and nutritious properties. In addition, mushrooms are also important in the prevention of diseases such as hypertension, Alzheimer's, Parkinson's and stroke risk, and are also used for other medical purposes [2]. Since mushrooms that can be grouped in two classes as edible and non-edible are easily accessible in nature, they are collected and consumed by the public especially in rural areas and as a result of this, poisoning cases are encountered. There is no single feature that distinguishes edible mushroom [3]. For this reason, some mushroom is consumed unconsciously and there are mild poisoning symptoms such as hallucinations, fever and nausea. The commercialization of mushroom has further enhanced the importance of detecting edible mushroom. Determination of the molecular and chemical properties of edible mushroom is both expensive and time consuming. However, some mushroom species can be distinguished more easily through some technologies and new statistical learning algorithms.

The classification algorithms mentioned in the scope of data mining are among the most used analyses. To use classification algorithms, classes must be known in advance. The class is divided into two groups: classes with predefined training set and test set. The model is created with the training set and the success of the model is checked with the test set. Algorithms with good model performance are used to determine the class of a new instance whose class is not known. In the literature, algorithms such as decision trees, support vector machines, Bayesian method and nearest neighbor are frequently used for classification analysis.

Cunningham and Holmes conducted classification studies with the Weka tool within the scope of developing innovative applications in agriculture and emphasized the support provided by the software tools. With this case study on mushroom grading, they discussed the data analysis process in the field of agriculture [4]. Alkronz et al. tried to predict whether mushroom is edible or non-edible by using back propagation artificial neural networks algorithm. At the end of the study, they achieved accurate estimation results with a rate of 99.25% [3]. Pinky et al. aimed to define the edibility of mushroom with the highest accuracy and the lowest error rate and to find the best community method [5]. Wibovo et al. showed that the performance of the C4.5 algorithm was better than the other two algorithms (naive Bayes, support vector machines) in their study to identify the edible mushroom and suggested that the results should be used in the generation of mobile applications [6]. Lidasay and Tagacay have developed a mobile-based application that combines the power of the neural network with image processing to identify the fungus by its image and whether it is edible or non-edible (poisonous) [7].

In the current study, the classification analysis which is one of the most important methods of data mining has been carried out based on the characteristics of the mushroom belonging to the agaricus and lepiota families which have edible and non-edible species. The characteristics of the mushrooms were examined with different decision tree algorithms (J48, RandomTree, RandomForest, LMT) with the help of Weka tool. The results were found to be parallel to other results in the literature and successful results were obtained with data mining analysis. In addition, in the current study, different from the literature, the fungal properties in the roots and nodes of the tree structures were examined and the features that play an important role in distinguishing the edible and non-edible mushroom were tried to be determined by decision tree algorithms.

2. MATERIAL AND METHOD

2.1 Data Set

Agaricus and lepiota mushroom samples were collected from the North American coast. This mushroom set is classified as edible, non-edible or unknown. However, in the scope of the study, the fungus group which was unknown was included in the non-edible class. The examples in the dataset have 23 attributes, including the class category. Of the 8124 observations, 2480 were recorded as missing data. Mushroom samples with missing data were not included in the analysis. The variables in the data set were measured as categorical data. Table-1 provides information about the data set.

Table 1: Information about the data set

Attribute Description	Attribute Type	Attribute Information
class	Nominal	e: edible, p: non-edible (poisonous)
cap-shape	Nominal	b: bell, c: conical, x: convex, f: flat, k: knobbed, s: sunken
cap-surface	Nominal	f: fibrous, g: grooves, y: scaly, s: smooth
cap-color	Nominal	n: brown, b: buff, c: cinnamon, g: gray, r: green, p: pink, u: purple, e: red, w: white, y: yellow
bruises	Nominal	t: bruises, f: no
odor	Nominal	a: almond, l: anise, c: creosote, y: fishy, f: foul, m: musty, n: none, p: pungent, s: spicy
gill-attachmment	Nominal	a: attached, d: descending, f: free, n: notched
gill-spacing	Nominal	c: close, w: crowded, d: distant
gill-size	Nominal	b: broad, n: narrow
gill-color	Nominal	k: black, n: brown, b: buff, h: chocolate, g: gray, r: green, o: orange, p: pink, u: purple, e: red, w: white, y: yellow
stalk-shape	Nominal	e: enlarging, t:tapering
stalk-root	Nominal	b: bulbous, c: club, u: cup, e: equal, z: rhizomorphs, r: rooted
stalk-surface-above-ring	Nominal	f: fibrous, y: scaly, k: silky, s: smooth
stalk-surface-below-ring	Nominal	f: fibrous, y: scaly, k: silky, s: smooth
stalk-color-above-ring	Nominal	n: brown, b: buff, c: cinnamon, g: gray, o: orange, p: pink, e: red, w: white, y: yellow
stalk-color-below-ring	Nominal	n: brown, b: buff, c: cinnamon, g: gray, o: orange, p: pink, e: red, w: white, y: yellow
veil-type	Nominal	p: partial, u: universal
veil-color	Nominal	n: brown, o: orange, w:white, y: yellow
ring-number	Nominal	n: none, o: one, t: two
ring-type	Nominal	c: cobwebby, e: evanescent, f: flaring, l: large, n: none, p: pendant, s: sheathing, z: zone
spore-print-color	Nominal	k: black, n: brown, b: buff, h: chocolate, r: green, o: orange, u: purple, w: white, y: yellow
population	Nominal	a: abundant, c: clustered, n: numerous, s: scattered, v: several, y: solitary
habitat	Nominal	g: grasses, l: leaves, m: meadows, p: paths, u: urban, w: waste, d: woods
number of instance: 5644 number of attribute: 23 missing value : no class distribution : edible: 3488 (%61.8) non-edible: 2156 (%38.2)		

In the current study, the characteristics of the roots and nodes in the tree structures were analyzed in order to determine the fungal properties affecting the classification results. Knowing the physical structure of mushroom is important in order to understand the fungal properties of the mushrooms analyzed in the study and to interpret the analysis results correctly. Therefore, the mushroom sections in Figure-1 [8] and a veiled agaricus mushroom sample [9] are shown in Figure-2.

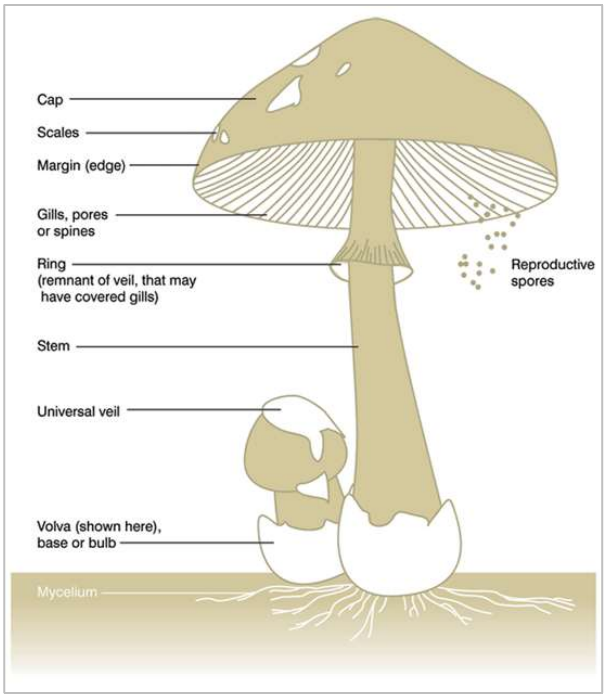

Fig.1. *Sections of mushroom.*

Fig.2. *An example of a partial veiled mushroom from the agaricus family.*

2.2 Weka Tool

It is an open source software developed in Java language at Waikato University. It is frequently preferred in data mining analyses [10,11]. It allows analysis by methods such as classification, clustering and association analysis. Not all file types are displayed on this interface. For this reason, the files need to be adapted to Weka. The most common file type is arff.

Weka is a very useful software for classification analysis. Once the data set is loaded, the classification analysis can be performed easily under the 'classify window'. Weka also has an option 'test options' section to select training and test series. Here you can set how much of the main data set is selected for training or testing. In this study, some decision tree algorithms and 10-fold cross validation test option in Weka were analyzed.

3. RESULTS

Table-2 shows the results for logistic model tree (LMT), correctly classified samples, misclassified samples, kappa statistics, mean absolute error, root mean square error, relative absolute error and root relative square error. Based on this evaluation in the training set, it is said that the LMT algorithm performs in the mushroom dataset with a 100% success rate. The run time of the algorithm was found to be 3.28 seconds. Since the number of trees and leaves is 1, the tree structure of the algorithm could not be obtained visually. However, the main feature that distinguishes the edible and non-edible is "odor" and then gill-size, stalk-shape, stalk-root, ring-number, spore-print-color, and population respectively.

Table 2: Performance of the mushroom dataset using Logistic Model Tree (LMT) algorithm

Evaluation on Training Dataset	**LMT**
Correctly Classified Instances	5644
Incorrectly Classified Instances	0
Kappa statistic	1
Mean absolute error	0.0268
Root mean squared error	0.0564
Relative absolute error	5.6735
Root relative squared error	11.6111
Time taken : 3.28 seconds Number of leaves : 1 Number of tree : 1	

According to Table 3, the correct classification of the J48 algorithm for mushroom data was 100% and the wrong classification was 0%. However, the average absolute error, root mean square error, relative absolute error and root relative square error perform well in the criteria. In terms of time, it is more economical than LMT with 0.01 second.

Table 3: Performance of the mushroom dataset using J48 algorithm

Evaluation on Training Dataset	**J48**
Correctly Classified Instances	5644
Incorrectly Classified Instances	0
Kappa statistic	1
Mean absolute error	0
Root mean squared error	0
Relative absolute error	0
Root relative squared error	0

Time taken : 0.01 seconds Number of leaves: 19 Number of tree : 25

Figure 3 is for decision tree structure of J48 algorithm result. When the structure of the tree was examined, as in LMT results, the "odor" variable was found as the root node of the tree. Ring number, veil color, spore print color, gill size, bruises were recorded as other distinctive characteristics.

```
odor = a: e (400.0)
odor = c: p (192.0)
odor = f: p (1584.0)
odor = l: e (400.0)
odor = m: p (36.0)
odor = n
|  ring-number = n: e (0.0)
|  ring-number = o
|  |  veil-color = w
|  |  |  gill-size = b: e (2496.0)
|  |  |  gill-size = n
|  |  |  |  bruises = f: e (144.0)
|  |  |  |  bruises = t: p (8.0)
|  |  veil-color = y: p (8.0)
|  ring-number = t
|  |  spore-print-color = h: p (0.0)
|  |  spore-print-color = k: p (0.0)
|  |  spore-print-color = n: p (0.0)
|  |  spore-print-color = r: p (72.0)
|  |  spore-print-color = u: p (0.0)
|  |  spore-print-color = w: e (48.0)
odor = p: p (256.0)
odor = s: e (0.0)
odor = y: e (0.0)
```

Fig.3. *J48 algorithm decision tree structure.*

Table 4 shows the performance of the mushroom data set based on the Random Tree classification method. When the results are examined, it is seen that 2 data are misclassified. Nevertheless, the results are quite successful. Algorithm time saving is better than LMT and J48 (0 second).

Table 4: Performance of the mushroom dataset using Random Tree algorithm

Evaluation on Training Dataset	Random Tree
Correctly Classified Instances	5642
Incorrectly Classified Instances	2
Kappa statistic	0.9992
Mean absolute error	0.0003
Root mean squared error	0.0148
Relative absolute error	0.0577
Root relative squared error	3.0493
Time taken : 0 seconds	

Number of leaves: - Number of tree : 68

Figure 4 shows the decision tree structure of the Random Tree algorithm. According to the results, stalk-surface-above-ring is the main factor determining the classification; odor, cap-color, gill-size, stalk-root, bruises, stalk-surface-below-ring, habitat, gill-spacing, and cup-surface are other characteristics that determine classification.

```
stalk-surface-above-ring = f
|  odor = a : e (0/0)
|  odor = c : e (0/0)
|  odor = f : p (144/0)
|  odor = l : e (0/0)
|  odor = m : e (0/0)
|  odor = n : e (408/0)
|  odor = p : e (0/0)
|  odor = s : e (0/0)
|  odor = y : e (0/0)
stalk-surface-above-ring = k : p (1332/0)
stalk-surface-above-ring = s
|  cap-color = b : p (72/0)
|  cap-color = c : e (20/0)
|  cap-color = e : e (576/0)
|  cap-color = g
|  |  odor = a : e (0/0)
|  |  odor = c : p (64/0)
|  |  odor = f : p (48/0)
|  |  odor = l : e (0/0)
|  |  odor = m : e (0/0)
|  |  odor = n : e (760/0)
|  |  odor = p : e (0/0)
|  |  odor = s : e (0/0)
|  |  odor = y : e (0/0)
|  cap-color = n
|  |  stalk-root = b : e (596/0)
|  |  stalk-root = c : e (0/0)
|  |  stalk-root = e
|  |  |  stalk-surface-below-ring = f : e (64/0)
|  |  |  stalk-surface-below-ring = k : e (0/0)
|  |  |  stalk-surface-below-ring = s
|  |  |  |  bruises = f : e (112/0)
|  |  |  |  bruises = t : p (128/0)
|  |  |  stalk-surface-below-ring = y : e (0/0)
|  |  stalk-root = r : e (96/0)
|  cap-color = p
|  |  bruises = f : p (64/0)
|  |  bruises = t
|  |  |  habitat = d : e (0/0)
|  |  |  habitat = g : p (12/0)
|  |  |  habitat = l : e (0/0)
|  |  |  habitat = m : p (12/0)
|  |  |  habitat = p : e (8/0)
|  |  |  habitat = u : e (0/0)
|  cap-color = r : e (0/0)
|  cap-color = u : e (0/0)
|  cap-color = w
|  |  odor = a : e (152/0)
|  |  odor = c : p (64/0)
|  |  odor = f : p (48/0)
|  |  odor = l : e (152/0)
|  |  odor = m : e (0/0)
|  |  odor = n
|  |  |  gill-spacing = c : p (24/0)
|  |  |  gill-spacing = w
|  |  |  |  cap-surface = f : e (64/0)
|  |  |  |  cap-surface = g : p (4/0)
|  |  |  |  cap-surface = s : e (64/0)
|  |  |  |  cap-surface = y : p (4/0)
|  |  odor = p : p (128/0)
|  |  odor = s : e (0/0)
|  |  odor = y : e (0/0)
|  cap-color = y : e (400/0)
stalk-surface-above-ring = y
|  gill-size = b : e (16/0)
|  gill-size = n : p (8/0)
```

Fig.4. *Random Tree algorithm decision tree structure.*

The REP Tree algorithm in Table 5, as in other algorithms, gave very good results. It correctly classified all observations and among other criteria the results were quite satisfactory. It is also successful in terms of time with 0.06 seconds.

Table 5: Performance of the Mushroom Dataset Using REP Tree Algorithm

Evaluation on Training Dataset	**REP Tree**
Correctly Classified Instances	5644
Incorrectly Classified Instances	0
Kappa statistic	1
Mean absolute error	0
Root mean squared error	0
Relative absolute error	0
Root relative squared error	0
Time taken : 0.06 seconds Number of leaves: - Number of tree : 26	

Figure 5 illustrates the decision tree structure of the REP Tree algorithm. The main distinguishing feature is the 'odor' attribute; spore-print-color and cap-color attributes are other distinctive features. This result is in parallel with other algorithm results.

```
odor = a : e (266/0) [134/0]
odor = c : p (124/0) [68/0]
odor = f : p (1052/0) [532/0]
odor = l : e (260/0) [140/0]
odor = m : p (27/0) [9/0]
odor = n
|  spore-print-color = h : e (0/0) [0/0]
|  spore-print-color = k : e (861/0) [435/0]
|  spore-print-color = n : e (870/0) [426/0]
|  spore-print-color = r : p (50/0) [22/0]
|  spore-print-color = u : e (0/0) [0/0]
|  spore-print-color = w
|  |  cap-color = b : e (0/0) [0/0]
|  |  cap-color = c : e (21/0) [11/0]
|  |  cap-color = e : e (0/0) [0/0]
|  |  cap-color = g : e (5/0) [3/0]
|  |  cap-color = n : e (36/0) [12/0]
|  |  cap-color = p : e (6/0) [2/0]
|  |  cap-color = r : e (0/0) [0/0]
|  |  cap-color = u : e (0/0) [0/0]
|  |  cap-color = w : p (5/0) [3/0]
|  |  cap-color = y : p (4/0) [4/0]
odor = p : p (175/0) [81/0]
odor = s : e (0/0) [0/0]
odor = y : e (0/0) [0/0]
```

Fig.5. *REP Tree algorithm decision tree structure.*

The results of the Hoeffding algorithm given in Table 6 produced lower results than the other algorithms. However, these results can still be considered successful at the decision stage. According to the results, the class of 43 observations could not be estimated correctly. Other statistical results seem to be acceptable. The time performance of the algorithm was found to be the same as the REP Tree algorithm. The decision tree structure of the algorithm was found to be the only root. It was observed that 'odor' is the basic and unique feature that distinguishes edibility.

Table 6: Performance of the mushroom dataset using Hoeffding Tree algorithm

Evaluation on Training Dataset	**Hoeffding Tree**
Correctly Classified Instances	5601
Incorrectly Classified Instances	43
Kappa statistic	0.9839
Mean absolute error	0.008
Root mean squared error	0.0787
Relative absolute error	1.6996
Root relative squared error	16.1941
Time taken : 0.06 seconds Number of leaves: - Number of tree : -	

Based on the results of the 5 different algorithms described above, Table 7 was created for the features that distinguish whether mushrooms are edible or non-edible. The values {1, 2, 3, 4, 5} in the table indicate the root level of the attributes. According to the table, it is clear that the "odor" attribute plays an important role in the classification of mushroom. Apart from this variable, gill-size, stalk-shape, stalk-root, ring-number, spore-print-color, population, veil-color, bruises, cap-color, stalk-surface-above-ring, stalk-surface-below-ring, habitat, gill-spacing variables are other variables that play an important role in differentiating the edibility of mushroom. In addition, when the decision tree structures given in Table 7 and above are taken into consideration, it is seen that Random Tree algorithm performs more detailed analysis using more mushroom features compared to the other algorithms.

Table 7: Mushroom attributes in decision tree structures

	LMT	J48	Random Tree	REP Tree	Hoeffding
1	**odor**	**odor**	stalk-surface-above-ring	**odor**	**odor**
2	gill-size, stalk-shape, stalk-root, ring-number, spore-print-color, population	ring-number	**odor,** cap-color, gill-size	spore-print-color	-
3	-	veil-color, spore-print-color	**odor,** stalk-root, bruises	cap-color	-
4	-	gill-size	stalk-surface-below-ring, habitat, gill-spacing	-	-
5	-	bruises	bruises, cap-surface	-	-

Table-8 was created with the help of ‘odor’ and algorithm results. In the light of the information in the table, the following general conclusions can be drawn:

- If the smell of mushrooms is almond, anise, fishy or spicy, it is edible.
- Non-edible if the smell of mushrooms is foul.
- If the smell of mushrooms is musty and they are stalk-surface-above-ring fibrous or stalk-surface-above-ring smooth and cap-color gray or stalk-surface-above-ring scaly and cap-color white, then they are edible.

  Otherwise mushrooms with musty smell are non-edible.
- If mushrooms do not smell, other characteristics of them (ring-number, spore-print-color, gill-spacing, stalk-surface-above-ring, cap-color, etc.) should be considered to make a decision.
- It is not correct to make any comments if the smell of mushrooms is creosote or pungent.

Table 8: Decision results

	J48	Random Tree			REP Tree	Hoeffding
		stalk-surface-above-ring				
		fibrous	smooth	scaly		
	odor		cap-color: gray	cap-color: white	**odor**	**odor**
odor		**odor**	**odor**	**odor**		
a	edible	edible	edible	edible	edible	edible
c	non-edible	edible	non-edible	non-edible	non-edible	non-edible
f	non-edible	non-edible	non-edible	non-edible	non-edible	non-edible
l	edible	edible	edible	edible	edible	edible
m	non-edible	edible	edible	edible	non-edible	non-edible
n	ring-number	edible	edible	gill-spacing	spore-print-color	edible
p	non-edible	edible	edible	non-edible	non-edible	non-edible
s	edible	edible	edible	edible	edible	-
y	edible	edible	edible	edible	edible	-

4. DISCUSSION

In the study, decision tree classification techniques performed extremely well in terms of accuracy (~ 100%). The results are strong for Kappa statistics, mean absolute error, root mean square error, relative absolute error and root relative square error.

This study can contribute to

- The development of image processing tools to identify edible and non-edible mushroom in the near future,
- The development of other methods for identifying factors that are effective in distinguishing mushrooms,
- The indication of the importance of using data mining methods in distinguishing other plants or animals,
- The prevention of unconscious mushroom picking, especially in rural areas.

In addition, it is known that laboratory investigations such as gene detection and chemical analysis are both costly and time consuming in the identification of fungal properties. In addition, determination of the main factors of mushroom by classical methods requires expertise. Therefore, classification methods may be preferred.

Considering the fact that mushrooms are collected and consumed in almost every region, the importance of identifying the characteristics that distinguish non-edible mushrooms is better understood. This study is expected to guide both readers and other researchers interested in mushroom.

5. References

[1] Bengü, A. Ş., Yılmaz, H. Ç., Türkekul, İ. & Işık, H. (2019). Doğadan toplanan ve kültürü yapılan pleurotus ostreatus ve agaricus bisporus mantarlarının toplam protein, vitamin ve yağ asidi içeriklerinin belirlenmesi. TTDBD, vol. 6(2), pp. 222-229.

[2] Husaini, M. (2018). A data mining based on ensemble classifier classification approach for edible mushroom identification. IRJET, vol. 5(7), pp. 1962-1966.

[3] Alkronz, E. S., Moghayer, K. A., Meimeh, M., Gazzaz, M., Abu-Nasser, B. S. & Abu-Naser, S. (2019). Prediction of whether mushroom is edible or poisonous using back-propagation neural network. IJAAR, vol. 3(2), pp. 1-8.

[4] Cunningham, S. J. & Holmes, G. (2001). Developing innovative applications in agriculture using data mining. Dept. Comput. Sci., Univercity Waikato, Hillcrest, New Zealand.

[5] Pinky, N. J., Islam, S. M. M. & Alice, R. S. (2019). Edibility detection of mushroom using ensemble methods. IJIGSP, vol. 4, pp. 55-62.

[6] Wibovo, A., Rahayu, Y., Riyanto, A. & Hidayatulloh, T. (2018). Classification algorithm for edible mushroom identification. 1sh ICOIACT, pp. 250-253. Yogyakarta, Indonesia.

[7] Lidasay, J. U. & Tagacay, M. P. (2018). Mushroom recognition using neural network. IJCS, vol. 15(5), pp. 52-57.

[8] The Mushroon Dairy (2019). Mushroom identification, UK Wild Mushroom Hunting Blog.

[9] Kummel, M. (2014). Agaricus mushroom shedding her partial veil.

[10] Ramya, R., Kumar Dr, P., Mugilan, D. & Babykala, M. (2018). A review of different classification techniques in machine learning using weka for plant disease detection. IRJET, vol. 5(5), pp. 3818-3823.

[11] Selvi, T. S. (2018). Efficient classification of agriculture land soils in statewise from india using data mining with Weka. IJSRCSEIT, vol. 3(3), pp. 44-51.