Application of Decision Tree Algorithm for Edible Mushroom Classification

Afika Rianti [1], Taufik Ridwan [2], Suprih Widodo [3], Rian Andrian [4]
Education Systems and Information Technology, Universitas Pendidikan Indonesia
afika@upi.edu [1], taufikridwan@upi.edu [2], supri@upi.edu [3], riananandrian@upi.edu [4]

Article Info

Article history:

Received 2022-05-20
Revised 2022-06-18
Accepted 2022-06-23

Keyword:

Classification,
Decision Tree Algorithm,
Edible Mushroom,
Overfitting,
Random Sampling Techniques.

ABSTRACT

This paper describes the classification of the mushroom based on its characteristic to be in an edible class or poisonous one using the Decision Tree Algorithm. The dataset was taken from the Kaggle website, while the tools used Weka. The result showed that odor is the most important attribute to classify the mushroom. Mushrooms which have almond and anise odors are edible, while the rest of it, such as pungent, foul, creosote, fishy, spicy, and musty are poisonous which means they can't be eaten. For mushrooms that have no odor, there are some attributes to be checked such as spore-print-color, gill-size, gill-spacing, and population. At first, overfitting happened. To overcome this, the researchers used Random Sampling Techniques until they got better accuracy. The result showed that this technique worked as it showed that 5/7 samples were not overfitting. The most accurate sample is 99,9% using sample 6 or 2000 data.

I. Introduction

Fungi affect human lives in many and varied ways like human diseases, plant pathogens, industrial products, human food (mushroom), and moreover that [1]. As human food, it means it's edible. Here people can use it as daily food to be eaten or just to make some processed food from mushrooms that they like. Even so, it doesn't mean people can eat all of the mushrooms. Mushrooms don't just have one kind. It has a lot from different genus and families. We could not eat any mushroom unless we were certain of its identity. Many mushroom species look alike and some species are highly poisonous. For that reason, people have to be careful. Because not all mushrooms are edible. Some mushrooms are not and they can be called poisonous ones.

Based on what is written on the news, there are some cases where people died because they ate mushrooms. For example, first, there's a case where a second child of an Afghan family that was evacuated from Kabul to Poland died after eating soup containing death cap mushrooms, which the family had unknowingly gathered in a Polish forest outside their quarantine center [2]. Then there's another case where a woman who is an ASHA worker died while 3 other members of the family were admitted to District Hospital, Senapati after ingesting the mushroom [3]. On the other hand, six people were killed in a remote village of West Jaintia Hills in Meghalaya after eating mushrooms [4]. Those cases are not the only cases that were reported. There are more than that.

In the year 1998 in France 1,675 cases of intoxication by mushrooms were reported and it is estimated that 8-10.000 cases are expected to be registered every year [5]. Said that most of them happened due to incorrect identification of species that are made by empirical and traditional knowledge [5]. Incorrect identification of mushrooms can be a big problem because if people carry on eating without knowing the truth that's safe or not, it can lead to death. So for sure, it will increase the number of cases where people die because of mushrooms.

Then also as there are many types of mushrooms so not everyone can remember. There's a need for something which can classify it using modern technology without needing to know which name, family, and type is the mushroom. Based on the problem above, the researchers have a solution to overcome this by classifying the mushroom-based on its characteristic to be in an edible or poisonous one using data mining techniques. Data mining is the discovery of interesting, unexpected, or valuable structures in large datasets [6]. Data mining itself has some algorithms such as Naive Bayes [7], Decision Tree [8], Random Forest [9], KNN [10], K-Means [11][12], etc. Each algorithm has a difference and so do the advantages and disadvantages. Here, the researchers used the Decision Tree algorithm to classify the mushroom classes. Decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable [13].

As written above, the researchers hope that this algorithm can classify the mushroom well based on its character and can give clues about how edible mushroom is and if it is not.

II. RESEARCH METHODOLOGY

A. Data

The data is a dataset that was collected from Kaggle website. Kaggle is an online community of data scientists and machine learners, owned by Google LLC [14]. This website gives freedom to people who want to use its dataset. The dataset of this research could be accessed on: https://www.kaggle.com/uciml/mushroom-classification [15]. The amount of this dataset is 8124 which it's about mushroom classification. Each data described a mushroom character from its looks and detail. This dataset contains 23 columns in total including the class of classification and has 8124 rows that show each data. Those 23 columns are class, cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color, population, and habitat.

B. Tools

The application that is used for classifying this dataset is Weka. The Waikato Environment for Knowledge Analysis (Weka) is a comprehensive suite of Java class libraries that implement many state-of-the-art machine learning and data mining algorithms [16]. The version of the Weka application that is used by the researchers is Weka 3.9. The researchers chose Weka because it's easy to use and it can interpret the result. Moreover, it doesn't lag too much and doesn't require big spaces.

C. Algorithm

The research methodology that is used in this data mining research is a Decision Tree. Decision trees work by grouping data one by one based on the value of each feature until the data is entered into a class. The class here is divided into two, namely poisonous and edible.

D. Procedure

The procedure of this research shown in Figure 1.

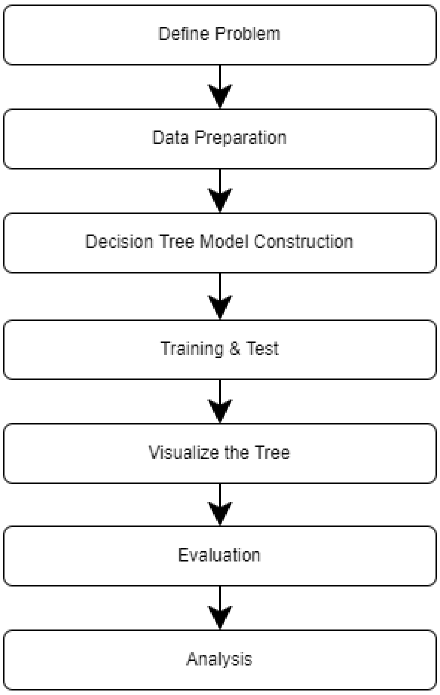


Figure 1. Research Procedure

1) Define Problem

The researchers first defined the problem and collected the related facts on introduction.

2) Data Preparation

The researchers then downloaded the dataset from Kaggle website then cleaned some data. For this research, the researchers only used 21 columns. Two columns were dropped. They are stalk root and veil types. Stalk root was dropped because it contains missing values. While veil-type was dropped because it has the same value for all data. As there's no difference between those data, the researchers would like to not use that column.

3) Decision Tree Model Construction

The researchers don't change the existing parameter values, so they still follow the defaults on Weka. The main parameters on the decision tree on Weka are maxDepth, noPrunning, numFolds, and minNum.

4) Training and Test

The mode test is split 66.0% train, remainder test.

5) Visualize the Tree

After getting the result, then we visualized the tree.

6) Evaluation

From the result, the researchers then do evaluation.

7) Analysis

As the researchers got what all needed from the research, the researchers did analysis, especially from the evaluation.

E. Evaluation Matrix

The evaluation metric used is the confusion matrix. Confusion Matrix is a method used to perform accuracy calculations on the concept of data mining [17]. The values displayed here are True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). True Positive means positive data that is predicted to be true. While True Negative means negative data that is predicted to be true. Then False Negative means negative data but is predicted as a positive word. Finally, False Negative means positive data but is predicted to be positive data. Those values are shown in Figure 2.

	POSITIVE	NEGATIVE
POSITIVE	True Positive (TP)	False Positive (FP)
NEGATIVE	False Negative (FN)	True Negative (TN)

Figure 2. Evaluation Matrix

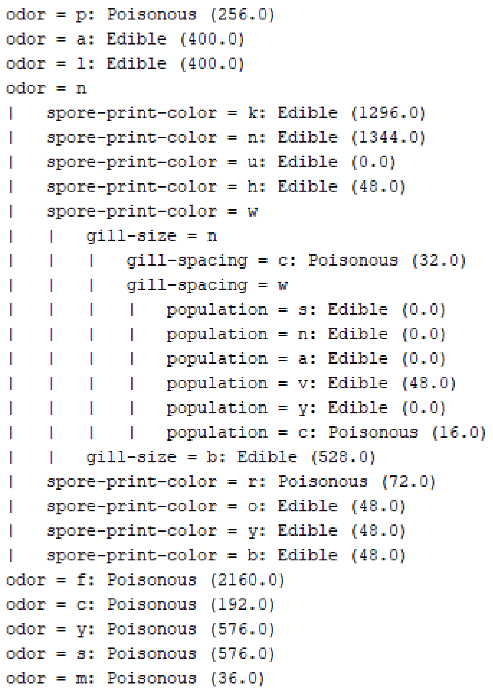
Actually, the confusion matrix not only performs accuracy only, but it can also perform other evaluations such as recall and precision. But for this research, the researchers chose to use accuracy only to be the parameter to determine how good the algoritma is. The accuracy formula is as follows.

$$\frac{TP + TN}{TP + FP + FN + TN}$$

We only need to add up the number of correct predictions per all data used as a test.

III. Result and Discussion

After the dataset was opened on Weka, the classification started. The mode test is split 66.0% train, remainder test. The result of the dataset that was used using Weka in Figure 3.



```
odor = p: Poisonous (256.0)
odor = a: Edible (400.0)
odor = l: Edible (400.0)
odor = n
|   spore-print-color = k: Edible (1296.0)
|   spore-print-color = n: Edible (1344.0)
|   spore-print-color = u: Edible (0.0)
|   spore-print-color = h: Edible (48.0)
|   spore-print-color = w
|   |   gill-size = n
|   |   |   gill-spacing = c: Poisonous (32.0)
|   |   |   gill-spacing = w
|   |   |   |   population = s: Edible (0.0)
|   |   |   |   population = n: Edible (0.0)
|   |   |   |   population = a: Edible (0.0)
|   |   |   |   population = v: Edible (48.0)
|   |   |   |   population = y: Edible (0.0)
|   |   |   |   population = c: Poisonous (16.0)
|   |   gill-size = b: Edible (528.0)
|   spore-print-color = r: Poisonous (72.0)
|   spore-print-color = o: Edible (48.0)
|   spore-print-color = y: Edible (48.0)
|   spore-print-color = b: Edible (48.0)
odor = f: Poisonous (2160.0)
odor = c: Poisonous (192.0)
odor = y: Poisonous (576.0)
odor = s: Poisonous (576.0)
odor = m: Poisonous (36.0)
```

Figure 3. Classification of Dataset

From the result above, we can see that odor is the most important attribute to classify the mushroom. Mushrooms which have almond(a) and anise(l) odor are edible, while the rest of it, such as pungent(p), foul(f), creosote(c), fishy(y), spicy(s), and musty(m) are poisonous which means they can't be eaten. For mushrooms that have no odor, there are some attributes to be checked such as spore-print-color, gill-size, gill-spacing, and population. Mushrooms which have no odor and have spore-print-color green(r) are poisonous. Then mushrooms which have no odor and have spore print-color black(k), brown(n), purple(u), chocolate(h), orange(o), yellow(y), and buff(b) are edible. And the mushroom which has no odor has white(w) spore print-color, and broad(b) gill-size is edible. Then for mushrooms that have no odor, white(w) spore-print-color, and close(c) gill-spacing is poisonous, same as the next mushroom with a crowded(w) gill-spacing and clustered(c) population. The rest of them are edible.

The visual tree to show how the dataset was classified based on its character is shown in Figure 4.

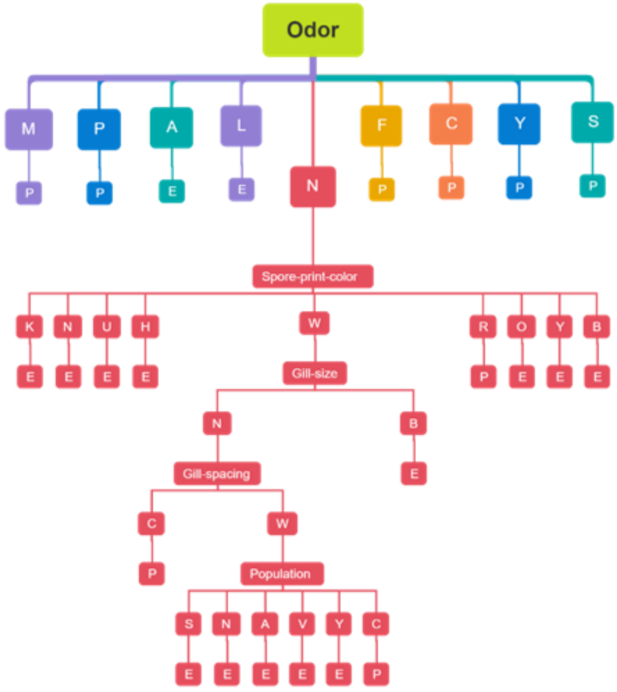
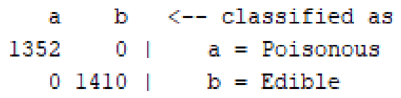


Figure 4. Visualization of Decision Tree

From that tree, we can get a point that mushrooms that have no odor, have a lot of attributes to be checked. Not like other odors which can be classified easily at the first root. The longer the tree, the greater attributes are used. Each value is only named with an initial such as M,P,A,L, etc.

Then for the evaluation, the confusion matrix and its accuracy for this case are shown in Figure 5.



```
    a    b   <-- classified as
 1352    0 |    a = Poisonous
    0 1410 |    b = Edible
```

Figure 5. Confusion matrix of a mushroom dataset

From those results, we can see that there are 1352 in true positive and 1410 in true negative. While false positive is 0 and false negative is 0. It means there's no error.

The accuracy for this case is as follows.

$$Accuracy = \frac{1352 + 1410}{2762} = \frac{2762}{2762} = 1 = 100\%$$

As we can see from the result above, the accuracy of this dataset using the decision tree algorithm is 100% which means it's overfitting. Overfitting is a fundamental issue in supervised machine learning which prevents us from perfectly generalizing the models to well fit observed data on training data, as well as unseen data on the testing set [17]. This could happen because too much data was used and some data from its character didn't make any difference so the result could be as the researchers expected.

There are some ways to handle overfitting such as reducing its effect by using various strategies like early-stopping, network reduction, data expansion, and regularization [18]. Another way to overcome this is using Random Sampling. The Random Sampling Technique (RST) has been previously used to improve the speed of a GP run [19]. In [20], Goncalves et al. (2012) showed that RST could decrease overfitting on all datasets.

Because of this reason, the researchers also used this technique to overcome the overfitting so that accuracy could be obtained. The research made seven different samples from its dataset. Each sample was taken randomly from its real dataset then was set to fit as some amount that was planned. For all of the modes, tests are split 66.0% train, remainder test, and so on for each sample. The result of these tests showed in table 1.

TABLE I
RANDOM SAMPLING

Sample	Total Data	Confusion Matrix				Accuracy
		TP	TN	FP	FN	
1	50	3	13	0	1	94.1%
2	100	10	23	1	0	97.1%
3	200	24	43	1	0	98.5%
4	500	83	87	0	0	100%
5	1000	146	192	2	0	99.4%
6	2000	348	331	1	0	99.9%
7	5000	602	1098	0	0	100%

From that table, we can see that overfitting still happened to sample 4 with 500 data and sample 7 with 5000 data. On the other hand, other samples are not overfitting such as 50, 100, 200, 1000, and 2000 data. As we can see, sample 6 has the most accuracy than the others. The accuracy on sample 6 is 99,9% then followed by sample 5 with 99,4% then sample 3 with 98,5% accuracy. From that table, we also can see the pattern about accuracy: the more data used, the higher accuracy will be. This happened to the 6/7 sample. While in sample 4, overfitting happened.

This accuracy may be different from the first one with 100% accuracy. Even that looks like a perfect number but in data mining, that shouldn't happen. Because when overfitting happens, it means there's something wrong. The accuracy shouldn't be so. The researchers chose one of the solutions to solve the problem using RST.

IV. Conclusion

The paper describes the classification of the mushroom based on its characteristic to be in an edible class or poisonous one using the Decision Tree Algorithm. The result on the first test using 8024 data is 100% accurate which means it's overfitting. To overcome this, the researchers did some tests again using Random Sampling Techniques where some data were taken randomly as some amounts to check its accuracy. This result showed that this technique worked as it showed that 5/7 samples were not overfitting. The most accurate sample is 99,9% using sample 6 or 2000 data, followed by sample 5 using 1000 data and sample 3 using 200 data. To know whether the mushroom is edible or not, people first have to check its odor. Mushrooms which have almond and anise odors are edible, while the rest of it, such as pungent, foul, creosote, fishy, spicy, and musty are poisonous which means they can't be eaten. For mushrooms that have no odor, there are some attributes to be checked such as spore-print-color, gill-size, gill-spacing, and population.

Acknowledgment

We would like to thank those who have helped us to complete this paper. Especially to Miss Nur Azizah who has given corrections to our paper by improving the arrangement of paper from the side of writing words.

References

[1] M. N. Owaid, M. M. Muslat, and W. C. Tan, ‘First collection and identification of wild mushrooms in western Iraq’, J. Adv. Lab. Res. Biol., vol. 5, no. 2, pp. 29–34, 2014.

[2] Anonim, “2nd Afghan Boy Dies of Mushroom Poisoning in Poland” Bloomberg.com, 03-Sep-2021. [Online]. Available: https://www.bloomberg.com/news/articles/2021-09-03/poland-probes-deadly-mushroom-poisoning-of-afghan-evacuees. [Accessed: 24-Dec-2021].

[3] Anonim, “Woman Dies after Consuming ‘Poisonous’ Mushroom” Thesangaiexpress.com/, 03-Sep-2021. [Online]. Available: https://www.thesangaiexpress.com/Encyc/2021/6/14/Our-CorrespondentSenapati-Jun-14-In-an-apparent-case-of-consuming-poisonous-mushroom-a-woman-w.html. [Accessed: 24-Dec-2021].

[4] Singh, Shipra, “How to Know a Mushroom is Edible or Poisonous?” Krishijagran.com, 03-Sep-2021. [Online]. Available: https://krishijagran.com/agripedia/how-to-know-a-mushroom-is-edible-or-poisonous. [Accessed: 24-Dec-2021].

[5] Lima, A & Fortes, Renata & Novaes, Maria Rita & Percario, Sandro. (2012). Poisonous mushrooms: a review of the most common intoxications. Nutrición hospitalaria : organo oficial de la Sociedad Española de Nutrición Parenteral y Enteral. 27. 402-8. 10.1590/S0212-16112012000200009.

[6] Hand, D. J. (2007). Principles of data mining. Drug safety, 30(7), 621-622.

[7] M. Y. Prasetyo, U. Darusalam, and B. Benrahman, ‘Web-Based Expert System for Diagnosis of Pigeon Disease by Naïve Bayes Method’, J. Appl. Inform. Comput., vol. 4, no. 2, Art. no. 2, Dec. 2020, doi: 10.30871/jaic.v4i2.2706.

[8] M. A. Hasanah, S. Soim, and A. S. Handayani, ‘Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir’, J. Appl. Inform. Comput. JAIC, vol. 5, no. 2, Art. no. 2, Oct. 2021, doi: 10.30871/jaic.v5i2.3200.

[9] B. Siswoyo, ‘MultiClass Decision Forest Machine Learning Artificial Intelligence’, J. Appl. Inform. Comput., vol. 4, no. 1, Art. no. 1, Jan. 2020, doi: 10.30871/jaic.v4i1.1155.

[10] P. Prasetyawan, I. Ahmad, R. I. Borman, Y. A. Pahlevi, and D. E. Kurniawan, ‘Classification of the Period Undergraduate Study Using Back-propagation Neural Network’, 2018, pp. 1–5.

[11] Y. Rokhayati, N. Z. Jannah, S. Irawan, and D. E. Kurniawan, ‘Decision Determination of Hinterland Selection Using Analytical Network Process’, in 2019 2nd International Conference on Applied Engineering (ICAE), Oct. 2019, pp. 1–5. doi: 10.1109/ICAE47758.2019.9221825.

[12] D. E. Kurniawan and A. Fatulloh, ‘Clustering of Social Conditions in Batam, Indonesia Using K-Means Algorithm and Geographic Information System’, Int. J. Earth Sci. Eng. IJEE, vol. 10, no. 5, pp. 1076–1080, 2017.

[13] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[14] Pratiwi, Banu Putri, Ade Silvia Handayani, and Bachelor Degree. "Measurement of Air Quality System Performance With Wsn Technology Using Confusion Matrix." Upgris Journal of Informatics 6, no. 2 (2020).

[15] Ni, H., Dong, X., Zheng, J., & Yu, G. (2021). An Introduction to Machine Learning in Quantitative Finance. World Scientific.

[16] Anonim, “Mushroom Classification” Kaggle.com, 02-Dec-2016. [Online]. Available: https://www.kaggle.com/uciml/mushroom-classification [Accessed: 24-Dec-2021].

[17] Holmes, G., & Hall, M. A. (2002). A development environment for predictive modelling in foods. International Journal of Food Microbiology, 73(2-3), 351-362.

[18] Ying, Xue. (2019). An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series. 1168. 022022. 10.1088/1742-6596/1168/2/022022.

[19] Gathercole, C., Ross, P.: Dynamic Training Subset Selection for Supervised Learning in Genetic Programming. In: Davidor, Y., M¨anner, R., Schwefel, H.-P. (eds.) PPSN 1994. LNCS, vol. 866, pp. 312–321. Springer, Heidelberg (1994)

[20] Gonçalves, Ivo & Silva, Sara & B. Melo, Joana & Carreiras, Joao. (2012). Random Sampling Technique for Overfitting Control in Genetic Programming. 218-229. 10.1007/978-3-642-29139-5_19.