Journal of Soft Computing Exploration

Accuracy of classification poisonous or edible of mushroom using naïve bayes and K-nearest neighbors

Roni Hamonangan[1], **Meidika Bagus Saputro**[2], **Cecep Bagus Surya Dinata Karta Atmaja**[3]
[12,3]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received Jan 8, 2021
Revised Feb 18, 2021
Accepted March 12, 2021

Keywords:

Accuracy
Mushroom Classification
Naïve Bayes
K-Nearest Neighbors

ABSTRACT

Mushrooms are plants that are widely consumed by the general public, but not all mushrooms can be consumed directly, because the types of mushrooms are feasible and it is still too difficult to distinguish, then there are several ways to identify fungi, namely by means of morphology. The morphology referred to in this paper is the morphology of fungi which includes color, habitat, class, and others. We got the morphology of this mushroom from a datasets we get from UCI Machine Learning with the 23 atribut that we use in the program. In determining the classification of this fungus we use the Naive Bayes algorithm which produces an accuracy of around 90,2% which we then improve again so that it reaches 100% accuracy using the K-Nearest Neighbors algorithm. Furthermore, in this case to prove accuracy that we had before, we use calculation accuracy with confusion matrix to show it the accuracy of classification poisonous or edible mushroom.

Corresponding Author:

Roni Hamonangan,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia.
Email: ronihamonangan93@students.unnes.ac.id

1. INTRODUCTION

Fungi belong to the fungal kingdom, therefore mushrooms do not have true leaves and roots, and do not have chlorophyll so they cannot carry out photosynthesis like plants in general. Fungi are classified or classified separately because they cannot be classified in plants or animals. There are fungi that can be seen directly or are macroscopic and some must be observed using a microscope or microscopic shape. In general, fungi have many cells (multicellular) such as edible mushrooms and tempeh mushrooms, but some are single-celled (unicellular) such as yeast or yeast (Saccharomyces). Multicellular fungi are composed of threads called hyphae. When viewed with a microscope, hyphae have a separating form (septa) and some are not partitioned [1].

A mushroom is one of the fungi types' food that has the most potent nutrients on the plant. Mushrooms have major advantages such as kill cancer cells, viruses and enhancing the human immune system. Currently, the mushroom refers to the process that performed by robot in food industry. This technique used to limit the features such as color. Recently, mushroom system used specific characteristics that improve the selection process of mushrooms. Such system depends on analyzing and investigating the features in order to get better classification based on the well-known features [2].

To identify which mushrooms are edible and poisonous, there are several ways that can be used. One of the aspects that can be used as benchmarks in identifying a fungus is its morphological characteristics. The morphological features referred to are the shape of the umbrella, color, habitat, and other features visible to our eyes. We obtained these morphological characteristics from the datasets we took from UCI Machine Learning [3].

Datasets is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set [4].

In this case we used two methods to determine the classification of mushrooms, namely the Naive Bayes method and also the K-Nearest Neighbor method as the classifier [5]. We use these two methods because they have different accuracy and we can compare it with the method which we get better accuracy of the two methods that have been tested. Extraction of morphological features is used to help identify fungi, so that later it will be known including the types of edible or poisonous mushrooms.

The accurate accuracy of the model can classify correctly [6]. Thus, the accuracy of the ratio of the predictions is correct (positive and negative) to the total data [7]. In other words, accuracy is the level of closeness of the predicted value to the actual (actual) value.

2. METHOD

In this study case the analysis will be carried out to find the best accuracy in determining the classification of fungi using two classification algorithms. The proposed algorithm is Naive Bayes algorithm [8] and K-Nearest Neighbor (KNN) algorithm [9], then evaluates and validates the results by looking for the best accuracy results of these two algorithms. The next stage is compare the results of the accuracy of each algorithm, to get a model classification algorithm that obtains the highest accuracy and time complexity [10]. The highest accuracy results from this calculation can be said to be the best algorithm in determining the classification of poisonous or edible mushrooms. And then we use calculation accuracy to test again the result of the accuracy that we get before.

For application in this case we will using Naive Bayes algorithm and K-Nearest Neighbor first [11]. Next we calculate using the Naive Bayes algorithm to get its accuracy, then we continue using the K-Nearest Neighbor algorithm to get better accuracy than the previous algorithm. The following is the flowchart that we produce according to the classification results of mushrooms using the Naive Bayes algorithm and the K-Nearest Neighbors algorithm, shown in Figure 1.

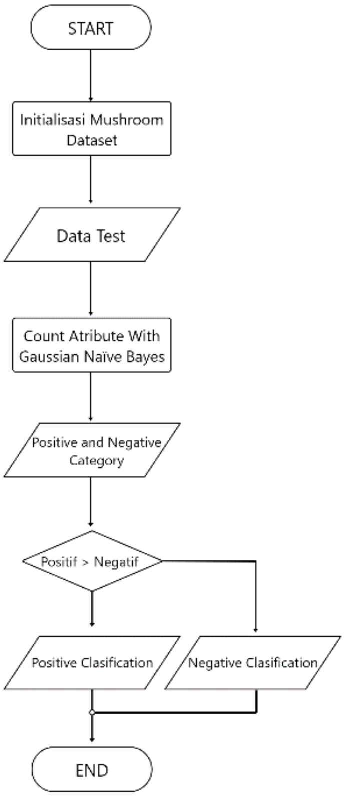

START
Initialisasi Mushroom Dataset
Data Test
Count Atribute With Gaussian Naïve Bayes
Positive and Negative Category
Positif > Negatif
Positive Clasification
Negative Clasification
END

Figure 1. Flowchart of naïve bayes algorithm

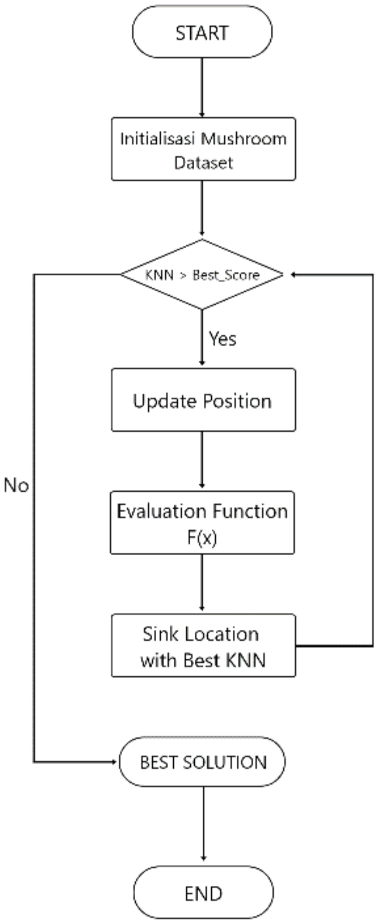
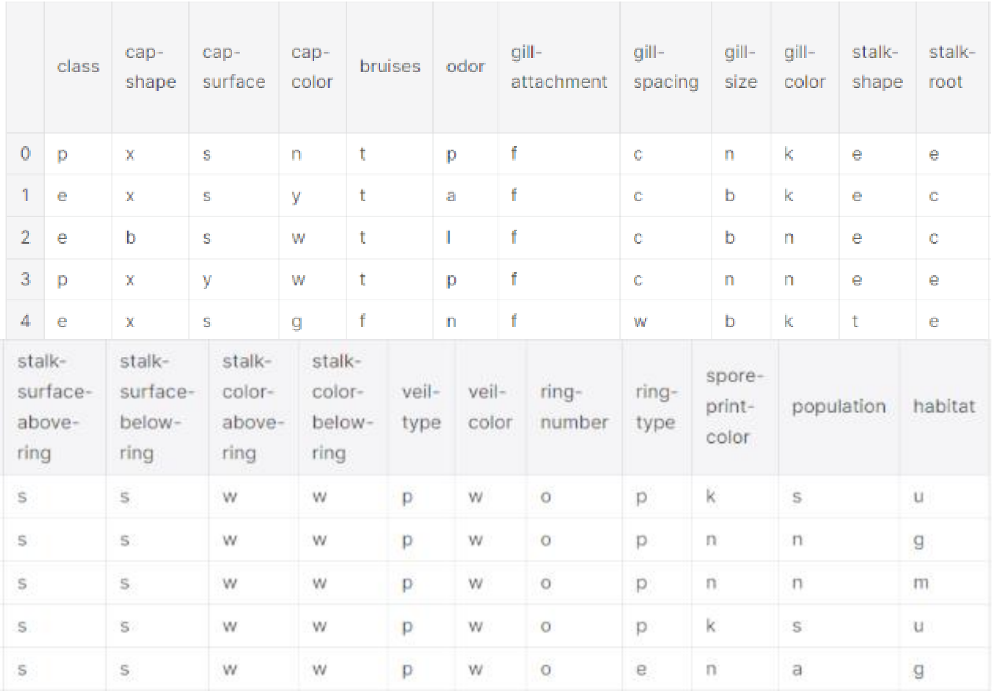


Figure 2. Flowchart of KNN algorithm

2.1 Data Collection

The data used is mushroom data obtained from the site archive.ics.uci.edu (UCI Machine Learning Repository) [12]. This data was donated by Jeffrey Schlimmer in 1987. In this study there are 2 classes, namely food mushrooms and poisonous mushrooms. For the number of each class, consisting of 4208 data included in the food mushroom category and 3916 data included in the poisonous mushroom category, so that the total number of data used was 8124 data. Each initial of each attribute and class is a representation of the type of attribute concerned. The feature extraction used in the training data is morphological features. The morphologicals that we use to calculate the accuracy of this classifocation poisonous or edible mushroom can show in the dataset that we have form UCI Machine Learning shown by the following this figure 3.



	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	habitat
0	p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	s	u
1	e	x	s	y	t	a	f	c	b	k	e	c	s	s	w	w	p	w	o	p	n	n	g
2	e	b	s	w	t	l	f	c	b	n	e	c	s	s	w	w	p	w	o	p	n	n	m
3	p	x	y	w	t	p	f	c	n	n	e	e	s	s	w	w	p	w	o	p	k	s	u
4	e	x	s	g	f	n	f	w	b	k	t	e	s	s	w	w	p	w	o	e	n	a	g

Figure 3. Datasets mushroom classification

2.2 Data Processing

In this mushroom classification process data, we use two algorithms, namely naive bayes and k-nearest neighbors to get the best accuracy in this classification. Here we use 23 attributes according to the datasets we use [13]. The attributes that we used in datasets there are class, cap shape, cap surface, cap color, bruises, odor, gill attachment, gill spacing, gill size, gill color, stalk root, stalk surface above ring, stalk surface below ring, stalk surface above ring, stalk surface below ring, veil type, veil color, ring number, ring type, spore print color, population, and last habitat of mushroom.

2.3 The Algorithm Used

2.3.1. Naïve Bayes

Bayes' theorem is a statistical calculation by calculating the probability of the similarity of an existing old case on a case basis with a new case [5]. Bayes' theorem has a high degree of accuracy and good speed when applied to large databases. Naive Bayes is included in supervised learning, so that at the learning stage, initial data is needed in the form of training data to be able to make decisions. At the classification stage, the probability value of each class label that is available for the input will be calculated. The class label that has the greatest probability value will be used as the label for the input data class. Naive Bayes is the simplest calculation of the Bayes theorem, because it is able to reduce computational complexity to a simple multiplication of probability. In addition, the Naive Bayes algorithm is also able to handle data sets that have many attributes.

The application of Caesarean Section data set on Naïve Bayes algorithm process as follows:

1. Prepare mushroom classification data set.
2. Classifying using Naïve Bayes algorithm.
3. Count the number of classes or labels in the data set.
4. Count the number of cases on each class.
5. Multiply all the class variables.
6. Compare the results of each classes.
7. The following is the equation of the Naïve Bayes:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{1}$$

In wich:

X : data or tuple object (class C)

H:	: hypothesis
P(H|X)	: probability that hypothesis H is in condition
P(H)	: prior probability that the H hypothesis is valid (true)
P(X)	: prior probability of tuple X

2.3.2. K-Nearest Neighbors

K-Nearest Neighbor or often abbreviated as KNN is one of the algorithms used to classify objects based on learning data (training data) which is the closest distance to the object [14]. The purpose of the KNN algorithm is to classify new objects based on attributes and samples from training data.

KNN is a supervised learning algorithm, which means that this algorithm uses existing data and the output is known. KNN is widely used in data mining, pattern recognition, image processing, etc.

1. Specifies the parameter K as the number of neighbors closest to the new object.
2. Calculate the distance between new objects / data against all objects / data that have been trained.
3. Sort the results of these calculations.
4. Determine the closest neighbor based on the minimum distance to K.
5. Determine the category of the closest neighbor to the object / data.
6. Use majority category as new object / data classification.

2.4 Calculation Accuracy

The method used to determine the final accuracy oftests performed is the confusion matrix for the multi-class method. This method is used to perform system calculations with many prediction classes [7]. The difference from the multi-class confusion matrix with the ordinary confusion matrix is that the final results are calculated cumulative accuracy of the overall accuracy of all test data. Parameters of the accuracy are presented in Table 1.

Table 1.Confusion matrix

Classification	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

where:

True Negative(TN)	: if the prediction and actual results are negative
False Negative (FN)	: if the positive prediction results, and the actual results negative
False Positive (FP)	: if the negative prediction results, and the actual results positive
True Positive (TP)	: if the predictive and actual results are positive.Calculation of the total accuracy of the tests performed using the following formula.

$$\text{Total Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \qquad (2)$$

3. RESULT AND DISCUSSION

To make this mushroom classification we use a software that is Google Colabs using python language. In determining this accuracy, we used the dataset we took from UCI Machine Learning, which was donated by Jeff Schlimmer in 1987. In this dataset, there are 23 attributes to calculate the accuracy of mushroom classification and with a lot of data, 8124 data.

3.1 Classification Mushroom with Naïve Bayes Algorithm

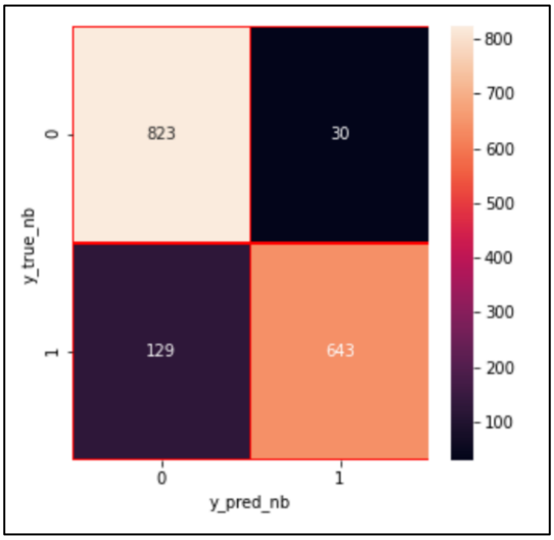


Figure 4. Confusion matrix of Accuracy Mushroom Classification with Naïve bayes

In Figure 3. It can be seen that the accuracy obtained using the Naive Bayes method is 90.21%. This accuracy is obtained from the import library in the form of GaussianNB to calculate the classification of this fungus. However, the calculation of the Naive Bayes method does not produce great accuracy results.

From the previous calculations that we have obtained before, we can prove it with the results we get the accuracy is 90,2%. To get the prove, one way to prove it is by using confusion matrix table. to show the confusion matrix table we can call the Naïve bayes values to the program and show the result to be output of this confusion matrix. The following is a calculation according to the confusion matrix that we can prove:

Accuracy =(TP+TN)/(TP+TN+FP+FN) × 100%
= (823+643)/(823+643+129+30) × 100%
=1466/1625 × 100%
= 0,9021 × 100%
= 90,2%

Calculation of accuracy in this study using a confusion matrix has been proven because we can see the result from true positive is 823 and true negative 643. At same time we can see if the value of false positive is 129 and false negative is 30. So, with that calculation we get the accuracy in this method using naïve bayes algorithm is 90,2%.

3.2 Classification Mushroom with K-Nearest Neighbors (KNN)

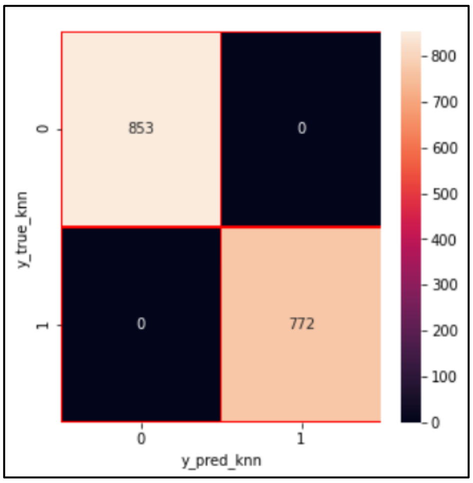


Figure 5. Confusion matrix of accuracy mushroom classification with KNN algorithm

Based on Figure 4, which calculates using the Naive Bayes method with an accuracy of 90.2% we think it has not reached the highest accuracy. Therefore, we tried to re-calculate the accuracy using the K-Nearest Neighbors algorithm and it is true that we found a very good accuracy of 100%. With using this KNN we need an import library, namely K-Nearest Neighbors Classifier. Not only that, we also calculated the best value from this classification of mushrooms by getting the best value, that is k = 1. So it can be concluded that by using the K-Nearest Neighbors algorithm we can produce a classification accuracy of this fungus reaching 100%.

From the previous calculations that we have obtained before, we can prove it with the results we get the accuracy is 100%. To get the prove, one way to prove it is by using confusion matrix table. to show the confusion matrix table we can call the KNN values to the program and show the result to be output of this confusion matrix. The following is a calculation according to the confusion matrix that we can prove:

$$\begin{aligned}\text{Accuracy} &= (TP+TN)/(TP+TN+FP+FN) \times 100\% \\ &= (853+772)/(853+772+0+0) \times 100\% \\ &= 1625/1625 \times 100\% \\ &= 1 \times 100\% \\ &= 100\%\end{aligned}$$

Calculation of accuracy in this study using a confusion matrix has been proven because we can see the result from true positive is 853 and true negative 772. At same time we can see if the value of false positive and false negative is 0. Therefore, to check the accuracy of mushroom classification with K-Nearest Neighbor algorithm is the best way to used that proven with 100% accuracy if we compare with Naïve Bayes algorithm.

4. CONCLUSION

The application of the two algorithms Naive Bayes algorithm and K-Nearest Neighbors produced two different accuracy as well. by using naive bayes algorithm based on our calculation, the results yield 90,2% accuracy. At the same time if we use K-Nearest Neihgbors based on the results of our calculations we produce an accuracy of up to 100%. So the best algorithm to determine the accuracy of the mushroom classification is the K-Nearest Neighbors algorithm.

REFERENCES

[1] H. G. Lewis and M. Brown, "A generalized confusion matrix for assessing area estimates from remotely sensed data," *Int. J. Remote Sens.*, vol. 22, no. 16, pp. 3223–3235, 2001, doi: 10.1080/01431160152558332.

[2] U. Lindequist, T. H. J. Niedermeyer, and W. D. Jülich, "The pharmacological potential of mushrooms," *Evidence-based Complement. Altern. Med.*, vol. 2, no. 3, pp. 285–299, 2005, doi: 10.1093/ecam/neh107.

[3] A. Srivastava, V. Singh, and G. S. Drall, "Sentiment analysis of twitter data: A hybrid approach," *Int. J. Healthc. Inf. Syst. Informatics*, vol. 14, no. 2, pp. 1–16, 2019, doi: 10.4018/IJHISI.2019040101.

[4] Kautsarina, A. N. Hidayanto, B. Anggorojati, Z. Abidin, and K. Phusavat, "Data modeling positive security behavior implementation among smart device users in Indonesia: A partial least squares structural equation modeling approach (PLS-SEM)," *Data Br.*, vol. 30, p. 105588, 2020, doi: 10.1016/j.dib.2020.105588.

[5] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, "Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, p. 427, 2018, doi: 10.25126/jtiik.201854773.

[6] G. Abdurrahman and J. T. Wijaya, "Analisis Klasifikasi Kelahiran Caesar Menggunakan Algoritma Naive Bayes," *JUSTINDO (Jurnal Sist. dan Teknol. Inf. Indones.*, vol. 4, no. 2, p. 46, 2019, doi: 10.32528/justindo.v4i2.2616.

[7] A. De Ramón Fernández, D. Ruiz Fernández, and M. T. Prieto Sánchez, "A decision support system for predicting the treatment of ectopic pregnancies," *Int. J. Med. Inform.*, vol. 129, pp. 198–204, 2019, doi: 10.1016/j.ijmedinf.2019.06.002.

[8] H. Manik, M. F. G. Siregar, R. Kintoko Rochadi, E. Sudaryati, I. Yustina, and R. S. Triyoga, "Maternal mortality classification for health promotive in Dairi using machine learning approach," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 851, no. 1, 2020, doi: 10.1088/1757-899X/851/1/012055.

[9] S. Sutarti, A. T. Putra, and E. Sugiharti, "Comparison of PCA and 2DPCA Accuracy with K-Nearest Neighbor Classification in Face Image Recognition," *Sci. J. Informatics*, vol. 6, no. 1, pp. 64–72, 2019, doi: 10.15294/sji.v6i1.18553.

[10] P. Radha and B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques," *.IJISET.* vol. 1, no. 6, pp. 334–339, 2014.

[11] N. R. Indraswari and Y. I. Kurniawan, “Aplikasi Prediksi Usia Kelahiran Dengan Metode Naive Bayes,” *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 9, no. 1, pp. 129–138, 2018, doi: 10.24176/simet.v9i1.1827.

[12] P. Ray and A. Chakrabarti, “Twitter sentiment analysis for product review using lexicon method,” *2017 Int. Conf. Data Manag. Anal. Innov. ICDMAI 2017*, pp. 211–216, 2017, doi: 10.1109/ICDMAI.2017.8073512.

[13] A. Al-Thubaity, Q. Alqahtani, and A. Aljandal, “Sentiment lexicon for sentiment analysis of Saudi dialect tweets,” *Procedia Comput. Sci.*, vol. 142, pp. 301–307, 2018, doi: 10.1016/j.procs.2018.10.494.

[14] K. Fukunaga and P. M. Narendra, “A Branch and Bound Algorithm for Computing k-Nearest Neighbors,” *IEEE Trans. Comput.*, vol. C–24, no. 7, pp. 750–753, 1975, doi: 10.1109/T-C.1975.224297.