Comparative Study of Machine Learning Techniques to Classify Edible and Poisonous Mushrooms

Sijimol Cyriac,
P G Scholar,
Amal Jyothi College of Engineering,
Kanjirapally, Kerala
sijimolcyriac@mca.ajce.in

Meera Rose Mathew,
Assistant Professor,
Amal Jyothi College of Engineering,
Kanjirapally, Kerala
meerarosemathew@amaljyothi.ac.in

***Abstract*—Mushrooms are members of the fungi kingdom, but they are classified as vegetables in cooking. Mushrooms come in a variety of shapes and sizes, and they can be both edible and poisonous. Each mushroom has its own appearance and flavour. However, the nutritional content of mushrooms varies depending on the type of mushroom used. Proteins, vitamins, minerals, amino acids, antibiotics, and antioxidants are among the essential nutrients found in them. Mushrooms are nutritionally beneficial to our bodies. However, not all mushroom species are edible; others are toxic, causing health issues and even death. As a result, it is necessary to verify if it is edible before consuming. The only way to eat mushrooms in a healthy manner is to determine and properly identify them. This paper compares the output of various machine learning techniques like OneR, k-Nearest Neighbors (KNN), J48, Random Forest on mushroom dataset in order to classify edible and poisonous mushrooms correctly.**

Keywords—Machine Learning, Mushrooms, OneR, k-Nearest Neighbors, J48, Random Forest, Weka, Accuracy

I. INTRODUCTION

Machine learning is a type of data analysis that automates the creation of analytical models. It's a field of artificial intelligence based on the premise that computers can learn from data, recognise patterns, and make decisions with minimal human intervention. This study aims to classify the mushroom dataset based on different features using the different techniques of Machine Learning (ML).

One of the most nutritious foods on the plant is a mushroom. Mushrooms have many benefits, including the ability to destroy cancer cells and viruses, as well as boosting the immune system of humans[1]. More research is being conducted on mushrooms every day because of their health benefits and different medicinal properties. Mushroom hunting, as the term implies, is the method of gathering mushrooms in the wild for food in most countries [2]. As a result, the general consensus on mushroom consumption is that only properly identified mushrooms can be consumed. We may not be able to distinguish between edible and non-edible mushrooms at this time. It is difficult for a non-expert to manually classify toxic and edible mushroom species of all species [2].To distinguish between poisonous and non-poisonous mushrooms, a computer-based system is needed [2]. Consumption of poisonous mushrooms by accident will result in death. As a result, it's important to distinguish between edible and poisonous mushrooms.

II. LITERATURE REVIEW

A Mushroom Diagnosis Assistance System (MDAS) was proposed by R. LaBarge[7], which consists of three components: a web application (server), a centralized database, and a mobile phone application (client) for use on mobile phones. The mushroom forms are determined using Naive Bays and Decision Tree classifiers. To begin, the suggested method selects the most well-known mushroom characteristics. Second, identify the type of mushroom. In terms of correct and incorrect graded cases, as well as error measurements, the Decision Tree classifier outperforms the Nave Bays classifier.

P. Babu[8], R. Thommandru[8], K. Swapna[8], and E. Nilima[8] proposed a new application domain that is used for SVM. For mushroom classification, the proposed method employs the Support Vector Machine and Nave Bayes algorithms. The results of the experiments revealed that SVM outperforms Nave Bayer's algorithm in terms of accuracy. Finally, the SVM is a powerful technique that can be applied to a variety of applications.

C.Eusebi[9], C.Gliga[9], D.John[9] and A. Maisonave[9] used various data mining techniques and the Weka mining method to evaluate a previous mushroom data collection. They used a voted perceptron algorithm, a nearest

DOI: 10.5281/zenodo.5091632
ISBN:978-93-5426-386-6@2021 *MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

neighbour classifier, a covering algorithm to collect correct rules, an unpruned decision tree, and a nearest neighbour classifier. After testing the techniques on various groups of stockholders, they discovered that an unpruned tree provides the highest accuracy and precision. Then it was used to create an interactive mushroom identification using a web-based human-machine programme.

S. Beniwal[10] and B. Das[10] used data mining classification techniques such as Zero, naïve Bayes and Bayes net to examine a mushroom dataset that included a variety of poisonous and nonpoisonous mushrooms. They used the precision, kappa statistic, and mean absolute error to test classification techniques. They discovered that the Bayes net has the smallest mean absolute error and the highest accuracy. and then naïve Bayes.

Shuhaida Ismail[3], Amy Rosshaida Zainal[3], Aida Mustapha[3] considered Mushroom features including shape, colour, and surface of the cap, stalk and gill, population and habitat, and odour. By sorting and rating the attributes, principal component analysis (PCA) is used to pick features. The Decision Tree (DT) is a classification algorithm in which each node represents a test of one or more attributes or features. The classification experiment is carried out using the Waikato Environment for Information Analysis (WEKA).Finally, the coefficient metric and processing times taken are used as assessment metrics to measure the established method classification accuracy.

III. METHODOLOGY

The aim of this study is to classify Mushroom dataset into two categories (poisonous and nonpoisonous) using machine learning techniques [1]. This section presents the classification methodology used to classify the mushrooms based on the behavioural features [3]. The methodology involves three important phases, which are collecting dataset, pre-processing and classification as shown in Figure 1.

A. WEKA

It is an open source software that includes tools for data pre-processing, machine learning algorithm implementation, and visualisation tools, allowing you to create machine learning techniques and apply them to real-world data mining problems [4].WEKA implements a number of algorithms in each of these categories. You can choose an algorithm and set the desired parameters and run it on the dataset [4]. Then, WEKA would give you the statistical output of the model processing. The various models can be applied on the same dataset [4]. The outputs of various models can then be compared, and the best model that matches your needs can be chosen. As a result, using WEKA speeds up the building of machine learning models in general [4].

B. Collecting Dataset

A number of variables are used to classify whether a mushroom is edible or poisonous. The dataset used in this study comes from the Kaggle repository, and it contains 8124 instances of mushrooms with 22 attributes and a known target class. This data set is used as training data for the classifier. Details of the variables that were used in the data are shown in Table I [3].

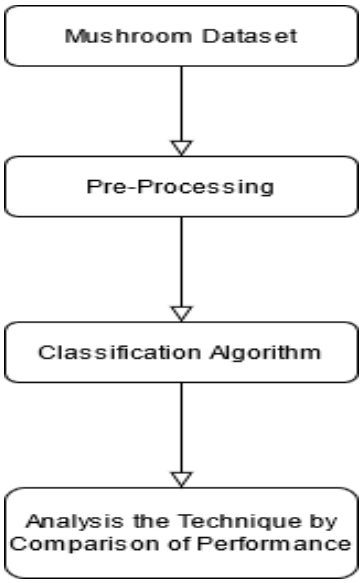


Fig. 1. Methodology for mushroom classification

C. Pre-Processing

In data mining, pre-processing is a crucial stage. It deals with entries that are missing, incomplete, or invalid for a variety of reasons, including data entry errors. In addition, there are also attributes that are not relevant to the research in data mining [3]. The irrelevant data such as ID should also be removed from the dataset because its presence can reduce the quality or accuracy of the data mining classification experiment [3]. The dataset for this study is made up of nominal values. These data must be turned into numerical values, which mean the nominal values must be converted into numerical values [3]. By transforming the nominal into numerical values, the data are now able to fed into classification algorithm [3].

DOI: 10.5281/zenodo.5091632
ISBN:978-93-5426-386-6@2021 *MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

D. Classification Experiment

The classification experiment is carried out using Waikato Environment for Knowledge Analysis (WEKA). In this paper, we will use different machine learning algorithms and techniques for mushroom classification such as J48, k-Nearest Neighbors, OneR, Random Forest and compare the results to see which algorithm has the best accuracy for the proper classification of poisonous and edible mushrooms.

E. Evaluation Matrix

An evaluation matrix is necessary for evaluating the results of a classification experiment [3]. The goal is to create a single measurement that can be used across multiple classification methods so that results can be compared [3]. In this work, the evaluation metrics used includes the classification accuracy, coefficient metric, and time taken to build the model.

TABLE I. FEATURES FOR THE MUSHROOM DATASET

Features	Nominal Values
Cap_shape	conical (*c*), convex (*x*), flat (*f*), bell (*b*), sunken (*s*), knobbed (*k*)
Cap_surface	grooves (*g*), scaly (*y*), smooth (*s*), fibrous (*f*)
Cap _colour	buff (*b*), cinnamon (*c*), grey (*g*), green (*r*), brown (*n*), purple (*u*), red (*e*), white (*w*), yellow (*y*), pink (*p*)
Bruises	no (*f*), bruises (*t*)
Odour	anise (*l*), creosote (*c*), fishy (*y*), foul (*f*), almond (*a*), none (*n*), pungent (*p*), spicy (*s*), musty (*m*)
Gill_attachment	descending (*d*), free (*f*), notched (*n*), attached (*a*)
Gill_spacing	crowded (*w*), distant (*d*), close (*c*)
Gill_size	narrow (*n*), broad (*b*)
Gill_color	brown (*n*), buff (*b*), chocolate (*h*), grey (*g*), black (*k*), orange (*o*), pink (*p*), purple (*u*), red (*e*), green (*r*), yellow (*y*), white (*w*)
Stalk_shape	tapering (*t*), enlarging (*e*)
Stalk_root	club (*c*), cup (*u*), equal (*e*), bulbous (*b*), rhizomorphs (*z*), rooted (*r*), missing (*?*)
Stalk_surface_above_ring	scaly (*y*), silky (*k*), smooth (*s*), fibrous (*f*)
Stalk_surface_below_ring	scaly (*y*), silky (*k*), smooth (*s*), fibrous (*f*)
Stalk_colour_above_ring	buff (*b*), cinnamon (*c*), grey (*g*), orange (*o*), brown (*n*), red (*e*), white (*w*), yellow (*y*), pink (*p*)
Stalk_colour_below_ring	buff (*b*), cinnamon (*c*), grey (*g*), orange (*o*), brown (*n*), red (*e*), white (*w*), yellow (*y*), pink (*p*)
Veil_type	universal (*u*), partial (*p*)
Veil_colour	orange (*o*), white (*w*), yellow (*y*), brown (*n*)
Ring_number	one (*o*), two (*t*), none (*n*)
Ring_type	evanescent (*e*), flaring (*f*), large (*l*), cobwebby (*c*), none (*n*), pendant (*p*), sheathing (*s*), zone (*z*)
Spore_print_colour	brown (*n*), buff (*b*), chocolate (*h*), green (*r*), black (*k*), purple (*u*), white (*w*), yellow (*y*), orange (*o*)
Population	clustered (*c*), numerous (*n*), abundant (*a*), several (*v*), solitary (*y*), scattered (*s*)
Habitat	grasses (*g*), leaves (*l*), meadows (*m*), paths (*p*), urban (*u*), waste (*w*), woods (*d*)

IV. IMPLIMENTATION

In this study, we will use different machine learning algorithms and techniques for mushroom classification and are listed below:

A. Classification Algorithms

J48 Algorithm: It is a decision tree technique and is one the most popular classification techniques in machine learning, where it used in decision support system. It allows you to predict a new dataset record's target variable [5].

The goal of using a J48 is to build a training model that can be used to predict the class or value of the target variable (training data) by learning simple decision rules derived from past data. When using Decision Trees to predict a record's class label, we start at the root of the tree. The values of the root attribute and the record's attribute are compared.

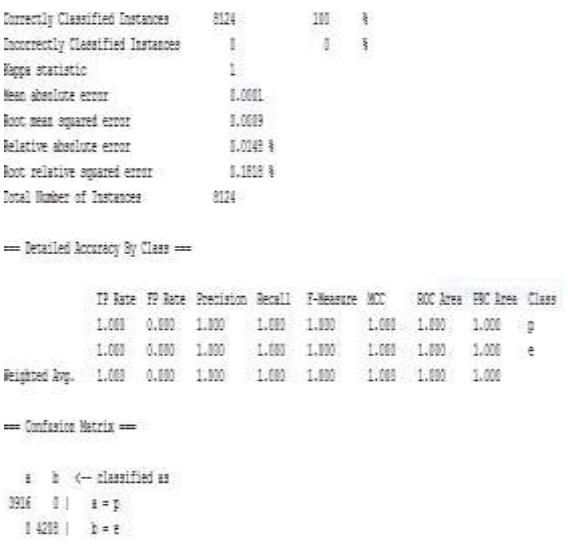
```
Correctly Classified Instances        8124               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances             8124

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1.000    0.000      1.000     1.000     1.000      1.000     1.000     1.000     p
                 1.000    0.000      1.000     1.000     1.000      1.000     1.000     1.000     e
Weighted Avg.    1.000    0.000      1.000     1.000     1.000      1.000     1.000     1.000

=== Confusion Matrix ===

    a    b   <-- classified as
 3916    0 |   a = p
    0 4208 |   b = e
```

Fig. 1. Classification accuracy by J48

Random Forest: It's a bagging extension of decision tree classification and regression. It creates a "forest" out of an ensemble of decision trees, which are normally trained using the "bagging" process.

The bagging method's basic premise is that combining several learning models improves the overall result [6]. The drawback of bagged decision trees is that they are constructed using a greedy algorithm that chooses the best split point at each level of the tree-building process [5].

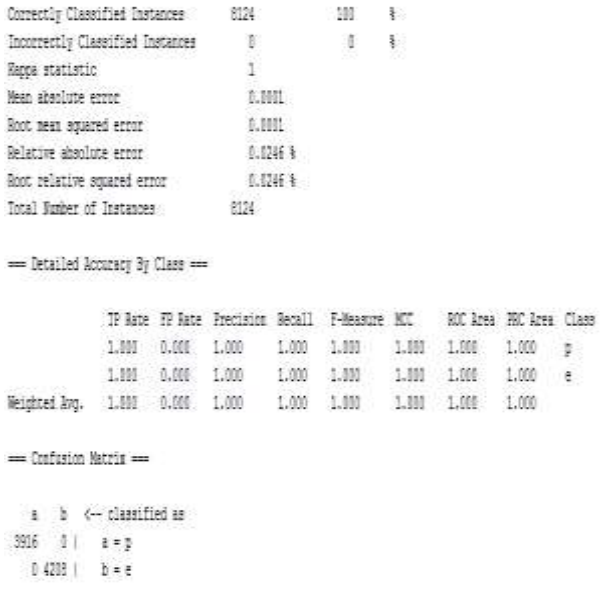
DOI: 10.5281/zenodo.5091632
ISBN:978-93-5426-386-6@2021 *MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*



```
Correctly Classified Instances        8124               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0001
Root mean squared error                  0.0009
Relative absolute error                  0.0049 %
Root relative squared error              0.1818 %
Total Number of Instances             8124

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 3916   0 |   a = p
    0 4208 |   b = e
```

Fig. 2. Classification accuracy by Random Forest

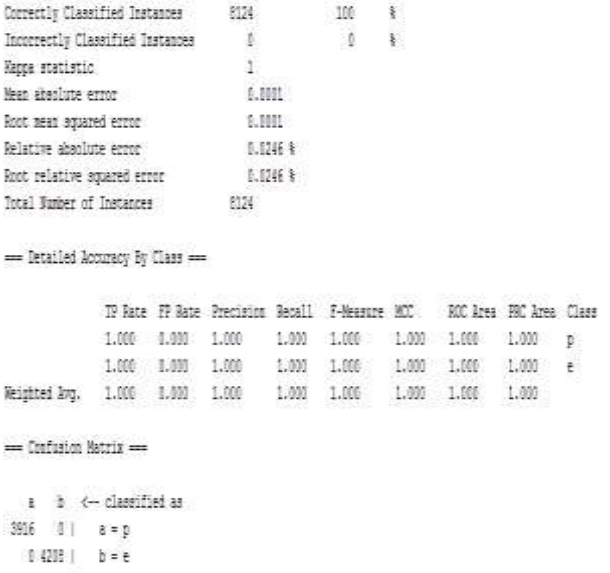
k–Nearest Neighbors: This approach is known as IBk in Weka (Instance Based Learner). Rather than building a model, the IBk algorithm generates a forecast for a test instance just-in-time. Using a distance metric, the IBk algorithm finds k "close" instances in the training data for each test instance, then generates a prediction based on those selected instances.



```
Correctly Classified Instances        8124               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0001
Root mean squared error                  0.0001
Relative absolute error                  0.0246 %
Root relative squared error              0.0246 %
Total Number of Instances             8124

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 3916   0 |   a = p
    0 4208 |   b = e
```

Fig. 3. Classification accuracy by k-Nearest Neighbors

OneR: For machine learning classification problems, the OneR classifier is one of the most simple and effective classifier algorithms. It creates one rule for each predictor in the data, and then chooses the rule with the smallest overall error to be its single rule.



```
Correctly Classified Instances        8124               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0001
Root mean squared error                  0.0001
Relative absolute error                  0.0246 %
Root relative squared error              0.0246 %
Total Number of Instances             8124

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     p
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     e
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 3916   0 |   a = p
    0 4208 |   b = e
```

Fig. 4. Classification accuracy by OneR

V. RESULT

The classification of this dataset was done to classify between edible and poisonous mushrooms based on their behavioural characteristics. The dataset had a total of 22 nominal attributes (features). Result of the testing using evaluation on the training dataset indicates that k-Nearest Neighbors has same accuracy level to J48 and Random Forest by 100%, and OneR has 98.5% accuracy. However, the k-Nearest Neighbors method is faster than J48 and Random Forest in terms of processing speed. Below Table1 shows the performance comparison chart of four algorithms on mushroom dataset.

TABLE 1. PERFORMANCE COMPARISON CHART

ML Techniques	**Accuracy**	**Time Taken to Build a Model**
J48 Algorithm	100 %	0.25s
Random Forest	100 %	2.96s
k-Nearest Neighbors	100 %	0.02s
OneR	98.5%	0.15s

DOI: 10.5281/zenodo.5091632
ISBN:978-93-5426-386-6@2021 *MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*

VI. CONCLUSION

Mushrooms have numerous advantages; however poisonous mushrooms can be lethal if consumed. It is so critical to correctly identify edible mushrooms before eating them. Here we consider mushroom dataset from Kaggle repository and applied different machine learning techniques to classify edible and poisonous mushrooms. This paper presented the methodology and results for mushroom classification experiment based on their behavioural features such as characteristics, population, and habitat. This comparison study has given us a better understanding of how the various machine learning models work and how they perform in real-world scenarios.

VII. FUTURE WORK

In data mining, we know that as the size of training dataset increases, the performance of the created model would be enhanced. As a future work, we can also include other attributes like cap margins, cap size, stem color, ecology, protein content, toxins, taste etc. and more records that are collected as part of research and apply different machine learning techniques on the new dataset. This will results in improving the model accuracy of the test.

VIII. REFERENCES

[1] Mohammad Ashraf Ottom, Noor Aldeen Alawad, Khalid M. O. Nahar "Classification of Mushroom Fungi Using Machine Learning Techniques"

[2] Rakesh Kumar Y, Dr. V. Chandrasekhar "Machine Learning Methods to Classify Mushrooms for Edibility-A Review"

[3] Shuhaida Ismail, Amy Rosshaida Zainal, Aida Mustapha "Behavioural Features of Mushroom Classification"

[4] https://www.tutorialspoint.com/weka/what_is_weka.htm

[5] N. Bhargava and G. Sharma, "Decision Tree Analysis on J48 Algorithm for Data Mining"

[6] https://builtin.com/data-science/random-forest-algorithm

[7] R. LaBarge, "Distinguishing Poisonous from Edible Wild Mushrooms"

[8] P. Babu, R. Thommandru, K. Swapna, and E. Nilima, "Development of Mushroom Expert System Based on SVM Classifier and Naive Bayes Classifier"

[9] C. Eusebi, C. Gliga, D. John, and A. Maisonave, "Data Mining on a Mushroom Database"

[10] S. Beniwal and B. Das, "Mushroom Classification Using Data Mining Techniques"

DOI: 10.5281/zenodo.5091632
ISBN:978-93-5426-386-6@2021 *MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam*