Machine Learning-Based Classification of Mushrooms for Edibility Assessment

Submitted by Group 2

1. INTRODUCTION

1.1. Background

Identification of mushrooms accurately as either edible or poisonous is a significant attempt due to the impact on human health and safety. Sri Lanka is home to a diverse array of edible wild mushrooms, many of which are exclusively consumed by indigenous communities with their traditional knowledge. However, misidentification of mushrooms can lead to severe health consequences, including organ failure and even death. With the help of Machine learning a solution can be invented to this critical issue. By utilizing algorithms and large datasets of mushroom characteristics, machine learning models can be used to differentiate mushrooms as edible and poisonous depending on the human obervations. By giving the physical attributes, like cap color, stem shape, and spore print, to analyze the data, those models can make highly accurate predictions depending on the given input data. This technology provides a rapid and reliable identification and allows individuals with limited mycological expertise to make decisions about mushroom edibility. This can reduce the potential risks associated with consumption of wild mushrooms. In summary, machine learning can play a vital role in enhancing the safety and well-being of humans by improving mushroom classification accurately.

1.2.Research Problem

Mushrooms are a very popular dish among the community. Tropical countries have high diversity of wild mushrooms rich in pharmaceutical and nutraceutical values. However, among them there are poisonous mushrooms, people get confused with their correct identification. To reduce the consuming life threatening consequences here aims to develop a machine learning model that can be used by any person without technical knowledge depending on the various physical characteristics (cap color, order, cap shape,etc.).

1.3. Data set Description

Data will be obtained from the Mushroom data set in the UCI Machine Learning Repository, which contains 22 different physical characteristics details of the 8124 samples from the Agaricus and Lepiota family. These variables will cover the features such as cap shape, cap surface, cap color, bruises, odor, gill attachment, gill spacing, gill size, 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, while the class label indicates the edibility of the mushrooms (edible or poisonous).

2. LITERATURE REVIEW

Mushrooms are the fruiting structures of microscopic fungi of Basidiomycetes, which grow naturally among the decaying woody substrates [1]. Generally, those mushrooms can be categorized in the major two types as edible and poisonous [2]. Among those species Family, Agaricus and Lepiota includes both edible and poisonous mushrooms, as well as being rich with pharmaceutical properties [3]. Machine learning base mushroom identifications have been conducted in different regions of the world with the help of Naïve Bayes and Voting Feature Interval(VFI5) algorithms. Those attempts had 99,552% and 84.53% prediction accuracy [4,5]
Naïve Bayes is one of classification algorithms and classification is one of the major studies in data mining [6]. Moreover, Rapid Miner or WEKA (Waikato Environment for Knowledge Analysis) was a popular data testing application among [2].

3.RESEARCH OBJECTIVE

Main objective:
To develop a machine learning model to preliminary identify the mushrooms either edible or poisonous based on their physical characteristics.
Specific Objectives:

- To collect data from US Irvine Machine Learning Repository with relevant physical characteristics
- To evaluate and select the suitable machine learning algorithms to gain the output results as binary (0/1); edible or poisonous mushroom.
- To analyze the significance of different physical characteristics (cap color, gil color, spore print color, odor etc.) to select the most indicative features in the edible mushrooms.
- To discuss the future potential to develop the model as a mobile application for assisting mushroom lovers to identify edible and poisonous mushrooms.

4. METHODOLOGY

4.1. Data Collection

Mushroom data set will be obtained from the USI Machine Learning Repository.

4.2. Data Processing

4.2.1. Data Cleaning

Handling missing values: Check for any missing values in the data set. Depending on the missing data amount, remove or fill those values.

4.2.2.Inserting categorical variable

Categorical variables will be inserted as numerical inputs. For this a numerical value will be assigned to each variable. In order to verify the results obtained, one-hot encoding methods also will be used to integrate the data into machine learning algorithms.

4.2.4.Splitting of Dataset

Whole dataset will be split into a training set and a testing set to evaluate the performance of the machine learning model with the help of Python.

4.2.3. Selection of features

Importance of all the characteristics which can be observed through the naked eye (physical) of the mushrooms (variables) will be analyzed to screen the indicative features of edible mushrooms, with the help of tree base models and correlation analysis.

4.2. Model Selection

A suitable machine learning algorithm will be used to evaluate the physical features of mushrooms. Basically, binary classification tasks like Decision Tree, Random Forest or Logistic Regression will be used with the consideration of factors like Data complexity, data set, accuracy, interpretability.

4.3. Model Training and Optimization

Different models will be trained with the data set, and finalize the best forming model. After that the selected machine learning model will fine tune to achieve high accuracy in the classification of mushroom data. To enhance the performance of the model, hyperparameter tuning and model optimization will be conducted.

4.4. Model evaluation and Validation

Performance of the model will be evaluated considering the evaluation metrics like, accuracy, precision,F1-score and recall.

4.5. Model Hyperparameter Tuning

Hyperparameters of the model will be fine tuned with the help of grid search, to increase its performance.

4.6. Visualization and Interpretation

Obtained results will be interpreted to understand the most significant physical characteristics to determine the edibility of the mushroom. Confusion matrices and ROC Curves (plots) will be used to interpret the results. These visualizations will be simple to give a better understanding to the non technical audience.

4.7. Compare with currently existing approaches

Final output will be compared with existing methods to observe the performance of the prepared model.

5.WORK PLAN

Milestones\Weeks	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
Proposal presentation										
Proposal Writing										
Research Methodology										
Final Report & Presentation										
Abstract submission for conference										

6. ACKNOWLEDGEMENT

Thanks to The Audubon Society Field Guide to North American Mushrooms which has provided data USI.

7.REFERENCES

[1] Rial Adity and Setia Hadi Purwono, Jamur – Info Lengkap dan Kiat Sukses Agribisnis. Depok, Indonesia/West Java: Agriflo, 2012.

[2] Wibowo, A., Rahayu, Y., Riyanto, A., & Hidayatulloh, T. Classification algorithm for edible mushroom identification. 2018 International Conference on Information and Communications Technology (ICOIACT).

[3] Bayu Mahardika Putra, "Klasifikasi Jamur ke Dalam Kelas Dapat Dikonsumsi Atau Beracun Menggunakan Algoritma VFI 5 (Studi Kasus : Famili Agaricus dan Lepiota)," IPB, Bogor, Laporan Akhir 2008.

[4] Bayu Mahardika Putra, "Klasifikasi Jamur ke Dalam Kelas Dapat Dikonsumsi Atau Beracun Menggunakan Algoritma VFI 5 (Studi Kasus : Famili Agaricus dan Lepiota)," IPB, Bogor, Laporan Akhir 2008.

[5] Galieh Adi and Surya Pradana, "Identifikasi jamur beracun pada jenis jamur famili agaricus dan lepiota berdasarkan klasifikasi," Univeritas 2018 International Conference on Information and Communications Technology (ICOIACT) 252 Nusantara PGRI Kediri, Kediri, Laporan Akhir 2016.