

Mushroom Poisonous Prediction Based on the Logistic Regression Model

Jian Sun*

College of Art and Science, New York University, New York, USA

*Corresponding author: js12359@nyu.edu

Abstract. Mushroom poisoning is a critical food safety issue that poses a significant threat to public health. In China alone, in 2022, there were 1332 reported cases of mushroom poisoning, leading to 28 deaths. With the existence of numerous toxic mushroom species that can cause fatal consequences, it is crucial to develop a reliable model for predicting mushroom toxicity. In this study, a logistic regression model was developed using a dataset consisting of over 60,000 records with 20 different variables. The model was constructed using the Python programming language and the Scikit-learn package. Logistic regression is a binary classification algorithm that uses a sigmoid function to transform its results into probability values. It relies on relationships between variables to predict a value and subsequently converts it to either a positive or negative outcome, corresponding to two classes. The results of this study indicated that the developed model achieved 100% prediction accuracy with the use of 2000 or more records in the dataset. Therefore, the proposed logistic regression model presents a promising tool for accurately predicting mushroom toxicity and mitigating the risks associated with mushroom poisoning. Further research could focus on expanding the dataset to include additional variables, such as environmental and geographical factors, to improve the model's accuracy and applicability.

Keywords: Mushroom Poisonous, Machine Learning, Logistic Regression.

1. Introduction

The world is home to approximately 140,000 species of macro fungi, with an estimated 50,000 to 100,000 fatalities resulting from mushroom poisoning annually, primarily in Europe, the United States, Japan, China, Iran, and other regions [1]. Common poisonous mushroom species are mostly distributed in some species e.g., Amanita genus, Cyclic mushroom [2-3]. Some mushrooms contain less toxic compounds and, therefore, are not severely poisonous. Poisonings by these mushrooms may respond well to treatment. However, certain types of mushrooms contain very potent toxins and are very poisonous; so even if symptoms are treated promptly, mortality is high. With some toxins, death can occur in a week or a few days. Although a liver or kidney transplant may save some patients with complete organ failure, in many cases there are no organs available. Patients hospitalized and given aggressive support therapy almost immediately after ingestion of amanitin-containing mushrooms have a mortality rate of only 10%, whereas those admitted 60 or more hours after ingestion have a 50–90% mortality rate.

In the United States alone each year, an average of seven people die from mushroom poisoning. Over 1,300 emergency room visits in the United States were attributed to poisonous mushroom ingestion in 2016, with about 9% of patients experiencing a serious adverse outcome. The North American Fungal Association (NAMA) Annual Report on Mushroom Poisoning showed a total of 690 cases from 2007 to 2014, with a mortality rate of 2.17 percent. A study in the United States (2020) reported an 8.8% death rate in 8953 mushroom poisoning cases [4]. In 2018 in Iran, following a string of mushroom poisoning events, more than 1200 cases were reported from 13 western and northwestern provinces, of whom 1.5% died [5].

Therefore, mushroom poisoning is a common and fatal factor in everyday life. Upon submission of mushroom specimens to laboratories, scientists may employ various methods, such as chemical tests and comparative analysis with prior data, to ascertain their toxicity. While chemical tests conducted in laboratories and hospitals are highly accurate, they often require a considerable amount of time. For most individuals, conventional wisdom dictates that brightly colored mushrooms should

not be consumed, leaving little else by way of guidance. Although some may attempt to identify mushrooms through online resources, they may still be confused by their previous experiences and some similar characters between poisonous and nonpoisonous mushrooms. Thus, there exists a need for a more rapid and precise approach to assess the safety of mushrooms found in one's vicinity.

To find a more effective way for identifying the type of mushroom, this study is intended to develop a mushroom classification method e.g., logistic regression by using many different characteristics including colors, shapes, smells, origins, etc.

2. Data-set description and preprocessing

This data set includes 61,069 hypothetical mushrooms with caps based on 173 species (353 mushrooms per species) [6]. Each mushroom is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended (the latter class was combined with the poisonous class). Of the 20 variables, 17 are nominal and 3 are metrical. The variables contain cap-diameter, cap-color, gill-color, stem-height, stem-weight, habitat, season, etc.

The present study involves the importation of a data set in CSV format into Jupyter hub followed by the removal of the header row containing variable names. A sub-sample consisting of 2000 rows is utilized for efficiency purposes. The data set is examined to ensure that no single category of either poisonous or edible mushrooms predominates. The resulting data set comprises 2000 rows and 21 columns, with the first column serving as the classification label for poisonous or edible mushrooms. The data set is partitioned into X and y subsets, with the former containing non-float instances. To facilitate the conversion of single characters to floats, OneHotEncoder is employed [7]. Ultimately, this prepared data set is utilized for logistic regression.

3. Machine learning model

Logistic regression is a binary classification algorithm based on linear regression that uses a sigmoid function to convert results into probability values. The goal of logistic regression is to predict which class (positive or negative) a given sample belongs to, based on its characteristics, with an output value of 0 or 1. To achieve this, a model through the relationship between sample features and labels should be built, that is, the parameters of the model are learned and updated by machine learning algorithms.

Specifically, this study first needs to pre-process and extract the features of the sample, and then enter the features and corresponding labels into the model. The parameter values in the model are continuously optimized with the training of the model, and finally the best model parameters are obtained, so as to obtain the optimal classification effect. The sigmoid function used in logistic regression algorithms is mainly used to map linear prediction results into the [0,1] interval and interpret the mapped results as probability values. In binary classification problems, the conventional approach is to classify samples having predicted values greater than or equal to 0.5 as positive classes, and those having predicted values less than 0.5 as negative classes.

In this experiment, the logistic regression method was implemented by sklearn package of python [8]. The whole data set is divided into X_train, X_test, y_train and y_test. Usually this study used X_train and y_train to train the model, where their rows correspond one by one. Each row of X is the input characteristics of one single record, and each row of y (actually one row of y only contains one result) is the right classification for corresponding record. After training, this model should be able to return the estimated probability of classification for each new record. Then import the X_test into this model, it will return a list of predicted result for each record in X. By comparing how much this predicted result is the same as y_train, the accuracy of this model can be obtained.

4. Results and Discussion

By using 2000 records from the data set, this model achieves a high level of success in predicting whether a specific mushroom is edible or poisonous, and by further testing, the accuracy still remains 98.7% after increasing the record number to 20000. For a higher number of training records, it is hard for this model to converge into a fixed number of iterations. Therefore, training datasets consisting of fewer than 8000 records are likely to be optimal for this model.

This model can also reveal how much each characteristic makes differences on predict results shown in Table 1, from the code below it can be observed that the coefficients and intercepts of models with 2000 and 3000 records. There are some coefficients always be zero, which means nearly no effects on results, so deleting these corresponding characteristics is reasonable. However, many of the coefficients are different while using different amounts of records, and the differences are dramatically large. This may be a problem about considering how much of the records should be used to train a proper model in this experiment.

Table 1. Coefficients with Different Training Dataset

Coefficients with 2000 training records	Coefficients with 3000 training records
-0.003	-0.004
-0.007	-0.005
-0.003	-0.002
0	0
-0.003	0
-0.006	-0.005
-0.006	-0.005
-0.007	-0.007
0	-0.005
-0.004	-0.002
-0.002	0
0	-0.002
-0.006	0
-0.010	-0.008
-0.006	-0.005
-0.012	-0.005
-0.006	-0.004
-0.004	-0.002
-0.007	-0.008
0.001	0.001

Finally, there is an issue that although many characteristics made zero effects on results can be eliminated, there are still more than 15 variables with similar coefficients (most coefficients are not really far away from each other). When people use this model, it is hard to collect accurate data for this high number of characteristics, and this may lead to the wrong results if only a few characteristics were collected. It is hard to solve this problem since most of the records contain some null values in different characteristics. In the future, more advanced models e.g., neural networks may be considered to obtain better performance in related tasks [9, 10].

5. Conclusion

This study utilized logistic regression as a classification technique for mushroom toxicity, achieving a 100% accuracy rate with a dataset of 20 variables. The model demonstrated the ability to learn classification using 2,000 to 10,000 records within a reasonable iteration time and produced clear classification probabilities approaching 0 or 1. However, one limitation of the model lies in its

dependence on 20 variables, which may be challenging to collect in practice. When only a few variables are available, the model's predictive capacity may be compromised, given that many of the coefficients are proximate to one another. Therefore, in further study, this model should be developed that could use less variables (delete less important variables) or distinguish a variable between null and not collected.

References

- [1] Diaz J H. Evolving global epidemiology, syndromic classification, general management, and prevention of unknown mushroom poisonings. *Crit Care Med*, 2005, 33(2): 419-426.
- [2] Schenk-Jaeger KM, Rauber-Lüthy C, Bodmer M, et al. Mushroom poisoning: A study on circumstances of exposure and patterns of toxicity. *Eur J Intern Med*, 2012, 23(4): e85-e91. DOI: 10.1016/j.ejim.2012.03.014
- [3] Brandenburg WE, Ward KJ. Mushroom poisoning epidemiology in the United States. *Mycologia*, 2018, 110(4): 637-641. DOI:10.1080/00275514.2018.1479561
- [4] De Olano J, Wang JJ, Villeneuve E, Gosselin S, Biary R, Su MK, et al. Current fatality rate of suspected cyclopeptide mushroom poisoning in the United States. *Clinical Toxicology*. 2020, 1–4.
- [5] Soltaninejad K. Outbreak of mushroom poisoning in Iran. April–May 2018 *The international journal of occupational and environmental medicine*. 2018, 9(3):152.
- [6] Kagglg, <https://www.kaggle.com/datasets/devzohaib/mushroom-edibility-classification>, 2022.
- [7] OneHotEncoder in sklearn.preprocessing, <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html>, 2023.
- [8] LaValley M P. Logistic regression. *Circulation*, 2008, 117(18): 2395-2399.
- [9] Yu Q, Yang Y, Lin Z, et al. Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV. *China Communications*, 2020, 17(3): 46-57.
- [10] Xie L, Wisse L E M, Wang J, et al. Deep label fusion: A generalizable hybrid multi-atlas and deep convolutional neural network for medical image segmentation. *Medical Image Analysis*, 2023, 83: 102683.