Mushroom classification by toxicity using logistic model

Alvin Gong

2020/12/13

Abstract

The research question is “Mushroom classification by toxicity” and it has some importance and application in real life situation. The study aims to find the association between mushroom traits and its toxicity. Using the cleaned data, we build a logistic regression model for predicting mushroom toxicity. Our results show that mushrooms with broad gills and pendant rings are more likely to be edible, while mushrooms with light spore-print color and bruises are more likely to be poisonous. In conclusion, there’s an association between mushroom traits and toxicity, but the causality is not yet determined. Therefore, further research need to be done to improve this model.

Keywords

Observational study, Mushroom toxicity, mushroom traits, estimated probability,log-odds,logistic regression model

Introduction

Mushrooms, which are fungi, are consumed for their nutritional and culinary value. However,not all mushrooms are edible, some are poisonous and they can lead to food poisoning or more severe symptoms(1). Thus,it’s very important to divide mushrooms into those can potentially be food and those cannot. Especially for people who love survival camping, it’s better not to eat mushrooms before making sure they’re not poisonous. There are around 5.1 million species of fungi(2), it’s not possible for us to know all species in detail right away when we come across to them, therefore it is essential to develop an easy way to identify which mushroom species are edible. This study will benefit survival camping lovers as well as many mycologists.

In this case, I will be performing analysis on mushroom toxicity using logistic model, I will investigate in the association between toxicity and traits(e.g.:spore-print-color,bruises,gill size,ring type,etc.). Once we know which traits come along with toxicity, we can apply this model to more mushrooms that we have less knowledge of, or even new species that we haven’t found before.

In the Data and Model section (Section 2), I describe the observational study, the data, and the logistic model created for mushroom classification. Results of the model analysis are provided in the Results section (Section 3),and discussions and conclusions are presented in Discussion section (Section 4), data citations and web citations are located in the References section (Section 5) at the end of report.

Data

The data that we use to do the analysis comes from Kaggle(3), which is a sample of various mushrooms with their corresponding toxicity. The data is collected through direct observations on different mushroom types found randomly, thus this is an observational study which includes no interference or manipulation.

Here is the attribute information of the cleaned dataset(**mushroom_data**):

Table 1 - Baseline characteristics of cleaned dataset

Variables	Description
bruises	1 indicates the mushroom can be bruised, 0 indicates it cannot
ring_type	cobwebby=c,evanescent=e,flaring=f,large=l,pendant=p,sheathing=s,zone=z
toxicity	1 indicates the mushroom is poisonous, 0 indicates it's edible
spore_light	1 indicates the mushroom has light-colored spore, 0 indicates it doesn't
spore_dark	1 indicates the mushroom has dark-colored spore, 0 indicates it doesn't
gill_broad	1 indicates the mushroom has broad gill, 0 indicates it has narrow gill

As described in *Table 1*, the cleaned data will include all above variables. **bruises** measures the mushroom's discoloration when squeezing it, it's a binary indicator(1=True or 0=False).**ring_type** is described using objects having similar shapes. **Toxicity** here is also binary, it is either poisonous or edible.**spore_light**,**spore_dark**,**gill_broad**, are dummy variables that work as their names suggest.

I selected some attributes based on Frederik Bussler's report(4), which uses the same dataset and featuring use of AI. In his report, he uploads the dataset to Apteo and automatically gets the variables that are indicative of a poisonous mushroom. After reviewing the results of machine learning, I selected some variables that have a "importance score" larger than 5, which means they are significant in predicting mushroom toxicity.

In the raw data, there are multiple colors under column **spore-print-color**, I combined black,brown and chocolate into the same category – spore_dark(dark-colored spore prints). And I combined buff, green,orange,purple, white, yellow into another category – spore_light(light-colored spore prints). Therefore, we have **6** variables and **8088** observations in the cleaned dataset.

Model

I will be using a logistics regression model to model the probability of a certain mushroom being poisonous. “Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary response variable” (5). The reason we choose this model is that the vote intention variable is binary, and “logistic regression is suitable when the outcome of interest is binary” (5).

I select the model based on AIC, a smaller AIC means better model. Using the **step** function, I found the model with the smallest AIC. Moreover,I choose this model because it contains most aspects that have more or less impact on mushroom **toxicity**.I want to see how all these variables can affect **toxicity**. Also, before seeing the results, I believe that lighter-colored and narrow-gill mushroom species will more likely to be poisonous since it might be a warning of danger to their predators.

I will be using bruises (dummy variable), ring_type (categorical variable), spore_light (dummy variable), spore_dark (dummy variable), gill_broad (dummy variable) to model the probability of mushroom being poisonous. The logistics regression model we are using is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{bruises1} + \beta_2 x_{ring\_typef} + \beta_3 x_{ring\_typel} + \beta_4 x_{ring\_typep} + \beta_5 x_{spore\_light} + \beta_6 x_{gill\_broad} + \epsilon$$

Where p represents the probability of mushroom being poisonous(toxicity=1). β_0 represents the intercept of the model, in this case, there is no practical interpretation for the intercept since every mushroom comes with certain traits, there’s no mushroom that has “zero trait”.

β_1 represents the log of odds ratio between mushrooms that can be bruised and mushrooms that cannot be bruised.β_2 represents if the mushroom has a flaring ring, we expect the log odds of the probability of the mushroom being poisonous to increase by a β_2, given other predictors hold constant.β_3 represents if the mushroom has a large ring, we expect the log odds of the probability of the mushroom being poisonous to increase by a β_3, given other predictors hold constant.β_4 represents if the mushroom has a pendant ring, we expect the log odds of the probability of the mushroom being poisonous to increase by a β_4, given other predictors hold constant.β_5 represents the log of odds ratio between mushrooms that have light spore-print colors and those do not have.β_6 represents the log of odds ratio between mushrooms that have broad gills and those do not have.

Result

Table 2 - Summary table of logistics regression model

Coefficients	Estimate	Std. error	z value	P-value
(Intercept)	1.4318	0.1350	10.606	< 2e-16
bruises1	1.1918	0.1263	9.433	< 2e-16
ring_typef	-20.9979	1552.2081	-0.014	0.98921
ring_typel	22.7875	298.7226	0.076	0.93919
ring_typep	-0.4053	0.1320	-3.071	0.00213
spore_light	1.1121	0.1238	8.985	< 2e-16
gill_broad	-4.6532	0.1189	-39.129	< 2e-16

From the *Table 2*, we can see that the estimated logistics regression model is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{bruises1} + \beta_2 x_{ring\_typef} + \beta_3 x_{ring\_typel} + \beta_4 x_{ring\_typep} + \beta_5 x_{spore\_light} + \beta_6 x_{gill\_broad} + \epsilon$$

$$= 1.4318 + 1.1918 x_{bruises1} - 20.9979 x_{ring\_typef} + 22.7875 x_{ring\_typel} - 0.4053 x_{ring\_typep} + 1.1121 x_{spore\_light} - 4.6532 x_{gill\_broad}$$

The $\hat{\beta}_0$ is the estimated **intercept** of the estimated logistic regression model, which is 1.4318. The standard error of $\hat{\beta}_0$ is 0.1350. The null hypothesis, H_0 is $\beta_0 = 0$, while the alternative hypothesis, H_a is $\beta_0 \neq 0$. Since the p-value $<$ 2e-16, which is much smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_1$ is the estimated coefficient of **bruise1(can be bruised)** of the estimated logistic regression model, which is 1.1918. The standard error of $\hat{\beta}_1$ is 0.1263. The null hypothesis, H_0 is $\beta_1 = 0$, while the alternative hypothesis, H_a is $\beta_1 \neq 0$. Since the p-value $<$ 2e-16, which is much smaller than 0.05, there is very strong evidence against the null hypothesis, H_0. It shows the difference in log odds of the probability of the mushroom being poisonous between the two groups(can be bruised and cannot be bruised) are most likely not due to random chance.

The $\hat{\beta}_2$ is the estimated coefficient of **ring_typef** of the estimated logistic regression model, which is -20.9979. The standard error of $\hat{\beta}_2$ is 1552.2081. The null hypothesis, H_0 is $\beta_2 = 0$, while the alternative hypothesis, H_a is $\beta_2 \neq 0$. Since the p-value is 0.98921, which is much larger than 0.05, there is no evidence against the null hypothesis, H_0. It shows the difference in log odds of the probability of the mushroom being poisonous between the two groups(have flaring ring and don't have flaring ring) may be due to random chance.

The $\hat{\beta}_3$ is the estimated coefficient of **ring_typel** of the estimated logistic regression model, which is 22.7875. The standard error of $\hat{\beta}_3$ is 298.7226. The null hypothesis, H_0 is $\beta_3 = 0$, while the alternative hypothesis, H_a is $\beta_3 \neq 0$. Since the p-value is 0.93919, which is much larger than 0.05, there is no evidence against the null hypothesis, H_0. It shows the difference in log odds of the probability of the mushroom being poisonous between the two groups(have large ring and don't have large ring) may be due to random chance.

The $\hat{\beta}_4$ is the estimated coefficient of **ring_typep** of the estimated logistic regression model, which is -0.4053. The standard error of $\hat{\beta}_4$ is 0.1320. The null hypothesis, H_0 is $\beta_4 = 0$, while the alternative hypothesis, H_a is $\beta_4 \neq 0$. Since the p-value is 0.00213, which is smaller than 0.05, there is strong evidence against the null hypothesis, H_0. It shows the difference in log odds of the probability of the mushroom being poisonous between the two groups(have pendant ring and don't have pendant ring) are more likely not due to random chance.

The $\hat{\beta}_5$ is the estimated coefficient of **spore_light** of the estimated logistic regression model, which is 1.1121. The standard error of $\hat{\beta}_5$ is 0.1238. The null hypothesis, H_0 is $\beta_5 = 0$, while the alternative hypothesis, H_a is $\beta_5 \neq 0$. Since the p-value $<$ 2e-16, which is much smaller than 0.05, there is very strong evidence against the null hypothesis, H_0. It shows the difference in log odds of the probability of the mushroom being poisonous between the two groups(having light spore-print color and don't have light spore-print color) are most likely not due to random chance.

The $\hat{\beta}_6$ is the estimated coefficient of **gill_broad** of the estimated logistic regression model, which is -4.6532. The standard error of $\hat{\beta}_6$ is 0.1189 . The null hypothesis, H_0 is $\beta_6 = 0$, while the alternative hypothesis, H_a is $\beta_6 \neq 0$. Since the p-value < 2e-16, which is much smaller than 0.05, there is very strong evidence against the null hypothesis H_0. It shows the difference in log odds of the probability of the mushroom being poisonous between the two groups(with broad gills and without broad gills) are most likely not due to random chance.

Based on the estimated logistics regression model considering the variable,we found out there's evidence showing that mushrooms that can be bruised and having light spore-print colors are more likely to be poisonous ; while mushrooms that having pendant ring-type and broad gills are more likely to be edible.

Then, we want to calculate numerically, by how much does the appearance of these traits add to the probability of a mushroom being poisonous. First we want to calculate the raw probability of a mushroom that has no trait being poisonous. Again, the value has no practical meaning since every mushroom comes with certain traits, just treat it as a base value for comparison purposes.

Since the estimated intercept is 1.4318, we get:

$$\hat{p}\_intercept = exp(1.4318)/(1 + exp(1.4318)) = 0.80718162169$$

Mushrooms that can be bruised:

$$\hat{p}\_bruise1 = exp(1.4318 + 1.1918)/(1 + exp(1.4318 + 1.1918)) = 0.93236507766$$

Mushrooms that have pendant ring:

$$\hat{p}\_ring\_typep = exp(1.4318 - 0.4053)/(1 + exp(1.4318 - 0.4053)) = 0.7362367854$$

Mushrooms that have light spore-print color:

$$\hat{p}\_spore\_light = exp(1.4318 + 1.1121)/(1 + exp(1.4318 + 1.1121)) = 0.92716264084$$

Mushrooms that have broad gills:

$$\hat{p}\_gill\_broad = exp(1.4318 - 4.6532)/(1 + exp(1.4318 - 4.6532)) = 0.03836829746$$

We don't include the cases when there's a flaring ring or large ring since the difference may just be due to random chance as stated above, there's no evidence for them to be significant when calculating probabilities.

Discussion

From analyzing mushroom toxicity data on Kaggle, we used logistic regression model to find out the association between mushroom traits and toxicity. Our model is based on **bruises**,**ring-type**,**spore-print color** and **gill size** , this means that these traits are most effective in predicting mushroom toxicity when we apply this model to more mushrooms.

From converting log-odds to probability, we can see that the appearance of bruises will increase the estimated probability of the mushroom being poisonous by approximately 12%, the appearance of pendant ring will decrease the estimated probability of the mushroom being poisonous by approximately 7%, the appearance of light spore-print color will increase the estimated probability of the mushroom being poisonous by approximately 12% and the appearance of broad gill will decrease the estimated probability of the mushroom being poisonous by approximately 77%, which is a huge difference here. Mushrooms that have broad gills are most likely to be edible since the estimated probability of a broad-gill mushroom being poisonous is an astounding 3.8%.

Mushroom hunting has a long history. It's still popular in many places today, especially in rural settings(4). From investigating my research question, we managed to develop a **guide** for survival camping lovers:broad-gill mushrooms are most likely to be edible, mushrooms can be bruised and have light spore-print colors are commonly poisonous, mushrooms that have pendant ring are more likely to be edible.However, this doesn't mean it's that certain. From 1999 to 2016, there were a whopping 133,700 cases of poisonous mushroom ingestion in just the United States. Sadly, 704 of these cases resulted in major harm, and 52 people died(6).The better way to avoid this is just **not to eat mushrooms living in the wild.**

Weakness & Next steps

Because observational studies are not randomized, they cannot control for all of the other inevitable, often immeasurable, exposures or factors that may actually be causing the results. Thus, any “link” between cause and effect in observational studies is speculative at best(7). In this report, I didn’t do any causal inference because the dataset is large, doing propensity score matching will increase the dimension, and make it too messy to draw conclusion from it. Thus, this study will not provide any insight on causality between toxicity and mushroom traits, the best we know is, there’s association between. The **“lurking variable”** in my guess is **species**. The dataset might include several mushrooms with different traits but they are the same species. Mutation happens on basically everything that has a DNA sequence, it might cause different individuals from same species to have different specialized traits, thus our conclusion drawn from only traits may not be correct. In my opinion, **species** is the real factor behind that determines toxicity.

Thus, to investigate this research question, we can use a different approach. If we only collect one mushroom with the least observable trait mutation from every species found, and put them together as a dataset, this eliminates the effect of **mutation and species** and we apply propensity score matching on it for **causality** purposes. This will provide us with a more precise and correct conclusion of this study.

References

(1) Mushroom Poisoning Syndromes. namyco.org/mushroom_poisoning_syndromes.php.

(2) M;, Blackwell. The Fungi: 1, 2, 3 … 5.1 Million Species? pubmed.ncbi.nlm.nih.gov/21613136/.

(3) Learning, UCI Machine. "Mushroom Classification." Kaggle, 1 Dec. 2016, www.kaggle.com/uciml/mushroom-classification.

(4) Bussler, Frederik. Deadly or Delightful-AI to Predict Mushroom Toxicity. 14 Aug. 2020, medium.com/towards-artificial-intelligence/deadly-or-delightful-ai-to-predict-mushroom-toxicity-2aaa81c98f7a.

(5) Caetano, S. (2020). Introduction to Logistic Regression. Lecture.

(6) authors, All, and William E. Brandenburg & Karlee J. Ward orcid.org/0000-0002-9222-7135. Mushroom Poisoning Epidemiology in the United States. www.tandfonline.com/doi/full/10.1080/00275514.2018.1479561.

(7) Observational Studies: Does the Language Fit the Evidence? Association vs. Causation. www.healthnewsreview.org/toolkit/tips-for-understanding-studies/does-the-language-fit-the-evidence-association-versus-causation/.