STA 223 Project 1: Analysis of Characteristics Influencing Mushroom Edibility

Christina De Cesaris

UC DAVIS UNIVERSITY OF CALIFORNIA

Introduction

- Estimated between 3 and 5 million species of fungi [1]
- Only 9.7% of recorded fungi form mushrooms [2]
- Most mushrooms are inedible
- It is possible that certain characteristics influence mushroom edibility status

Project Goal: Investigate the relationship between mushroom specie features and mushroom edibility to humans utilizing the profiles of 173 mushroom species found in the dataset [4]

Data Description and Preprocessing

Raw: 20 Predictors, 173 samples (3 numeric, 17 categorical) [5]

Processed: 70 Predictors, 173 samples (6 numeric, 64 categorical)

Step 1

Variable Name	Description
family	Taxomic Family
name	Species Name
class	Edibility Status
cap.diameter	Diameter of Mushroom Cap (cm)
cap.shape	Mushroom Cap Shape
Cap.surface	Cap Surface Texture
cap.color	Color of Cap
does.bruise.or.bleed	Does Mushroom Bruise or Bleed (t/f)
gill.attachment	Gill Attachment Present
gill.spacing	Gill Spacing Type
gill.color	Gill Color
stem.height	Stem Height (cm)
stem.width	stem.width (mm)
stem.root	Stem Root Type
stem.surface	Stem Surface Type
stem.color	Stem Color
veil.type	Veil Type
veil.color	Veil Color
has.ring	Is a Ring Present (t/f)
ring.type	Type of Ring
Spore.print.color	Color of Spores
habitat	Native Habitat
season	Growing Season

Table 1: Dataset Variables and Descriptions

Variable	Percent Missing
cap.diameter	0.000000
cap.shape	0.000000
Cap.surface	**23.121387**
cap.color	0.000000
does.bruise.or.bleed	0.000000
gill.attachment	**16.184971**
gill.spacing	**41.040462**
gill.color	0.000000
stem.height	0.000000
stem.width	0.000000
stem.root	**84.393064**
stem.surface	**62.427746**
stem.color	0.000000
veil.type	**94.797688**
veil.color	**87.861272**
has.ring	0.000000
ring.type	4.046243
Spore.print.color	**89.595376**
habitat	0.000000
season	0.000000

Table 2: Percent of missing values for predictors

class	cap.diameter	cap.shape	cap.color	gill.color	stem.height	stem.width	stem.color	has.ring	ring.type	habitat	season
p	[10, 20]	[x, f]	[e, o]	[w]	[15, 20]	[15, 20]	[w]	[t]	[g, p]	[d]	[u, a, w]
p	[5, 10]	[p, x]	[n]	[w]	[6, 10]	[10, 20]	[w]	[t]	[p]	[d]	[u, a]
p	[10, 15]	[x, f]	[g, n]	[w]	[10, 12]	[10, 20]	[w]	[t]	[e, g]	[d]	[u, a]

Table 3: Excerpt of Raw data after dropping missing variables

Step 2

Color	Color Code	Ring Type	Ring Code	Habitat Type	Habitat Code	Cap Shape	Cap Code	Season	Season Code
brown	n	cobwebby	c	grasses	g	bell	b	Spring	s
buff	b	evanescent	e	leaves	l	conical	c	Summer	u
gray	g	flaring	r	meadows	m	convex	x	Autumn	a
green	r	grooved	g	paths	p	flat	f	Winter	w
pink	p	large	l	heaths	h	sunken	s		
purple	u	pendant	p	urban	u	spherical	p		
red	e	sheathing	s	waste	w	others	o		
white	w	zone	z	woods	d				
yellow	y	scaly	y						
blue	l	movable	m						
orange	o	none	f						
black	k	unknown	?						

Table 4: Retained categorical predictor codes

cap.diameter	season	cap.diameter_max	cap.diameter_min	season_a	season_s	season_u	season_w
[10, 20]	[u, a, w]	**20**	**10**	**1**	**0**	**1**	**1**
[5, 10]	[u, a]	**10**	**5**	**1**	**0**	**1**	**0**
[10, 15]	[u, a]	**15**	**10**	**1**	**0**	**1**	**0**

Table 5: Data processing example with numerical and categorical variable

Methods

Binary outcome variable (inedible = 1) is suitable for analysis using a logistic regression model [3]

Model:

$$\text{Link Function: } \eta = log\left(\frac{p(y)}{1-p(y)}\right)$$

$$Logistic\ Regression: \eta = \beta_1 X_1 + \cdots + \beta_p X_p\ for\ p\ predictors$$

- Backwards selection was performed on the full model using AIC and BIC criteria

$$\eta_{BIC} = \beta_0 + \beta_1 season_{winter} + \beta_2 cap.shape_{bell} + \beta_3 cap.color_{brown} + \beta_4 cap.color_{green} + \beta_5 stem.color_{white} + \beta_6 ring.type_{zone}$$

- All predictors but $\boldsymbol{ring.type_{zone}}$ significant to the level 0.05.
- $\mathbf{ring.type_{zone}}$ ultimately dropped from the model

	Estimate	Std. Error	P Value	5% CI	95% CI
$\mathbf{ring.type_{zone}}$	-16.6113	869.5533	0.984759	NA	47.6069152

Table 6: Ring Type Zone model output

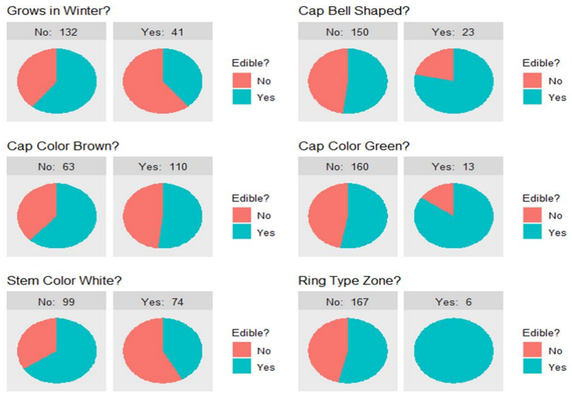

Figure 1: BIC Predictors plotted by edibility status

Results

$$\eta_{\text{final}} = \beta_0 + \beta_1 season_{winter} + \beta_2 cap.shape_{bell} + \beta_3 cap.color_{brown} + \beta_4 cap.color_{green} + \beta_5 stem.color_{white}$$

	Estimate	Std. Error	P Value	CI 5%	CI 95%
Intercept	1.3156	0.3915	0.000778 ***	0.6913500	1.9840120
$\mathbf{season_{winter}}$	-0.9822	0.4011	0.014345 *	-1.6573755	-0.3323078
$\mathbf{cap.shape_{bell}}$	1.5043	0.6010	0.012317 *	0.5676204	2.5649473
$\mathbf{cap.color_{brown}}$	-0.8634	0.3827	0.024068 *	-1.5098363	-0.2465492
$\mathbf{cap.color_{green}}$	1.8991	0.8395	0.023693 *	0.6493048	3.4973006
$\mathbf{stem.color_{white}}$	-1.3404	0.3688	0.000279 ***	-1.9639632	-0.7466615
Null deviance: 237.74 on 172 df			**Residual deviance:** 203.69 on 167 df		
AIC: **215.69**					

Table 7: Final model output

Diagnostics

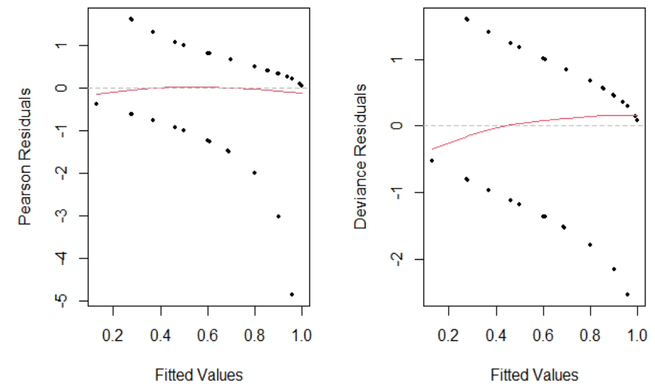

Figure 3: Residuals vs fitted plot for final model

Discussion

- Mushrooms with white stems, winter growing season, and brown caps have a lower probability of being inedible when compared to those without the respective variables
- Mushrooms with bell shaped caps, and green caps have a higher chance of being inedible compared to respective baselines
- White stem presence was the most significant predictor in the model and the most balanced
- Additional profiles could be included to improve balancing and possibly predictor significance
- A dataset consisting of live recorded samples may be more useful in practice
- Max-Min range integer variables not informative
- No severe lack of fit from diagnostics

References

[1] Blackwell M. (2011). The fungi: 1, 2, 3 ... 5.1 million species ?. *American journal of botany*, *98*(3), 426–438. https://doi.org/10.3732/ajb.1000298

[2] Casadevall, A., Heitman, J., & Buckley, M. (2008). The Fungal Kingdom: Diverse and Essential Roles in Earth's Ecosystem.

[3] Faraway, J. (2016). Extending the linear model with R. Second Edition, Chapman and Hall. ISBN 9781498720960

[4] Harding, P., (2013). *Mushroms and Toadstools*. Dorling Kindersley.

[5] Wagner, D., Heider, D., & Hattab, G. (2021). Mushroom data creation, curation, and simulation to support classification tasks. *Scientific reports*, *11*(1), 8134. https://doi.org/10.1038/s41598-021-87602-3