Project 2 Part A

● Graded

Group

Samantha Zheng
Autumn Xu
Yueping Gu
✎ View or edit group

Total Points

13 / 13 pts

Question 1

Data — **3** / 3 pts

1.1 **Dataset** — **1** / 1 pt

✓ **+ 1 pt** Correct

+ 0 pts Not provided/Insufficient

1.2 **Variables** — **2** / 2 pts

✓ **+ 2 pts** Correct

+ 1 pt Insufficient

+ 0 pts Left blank

Question 2

Define a real-world motivation

4 / 4 pts

✓ **+ 4 pts** Correct

+ 2 pts Partial

+ 0 pts Left blank

Question 3

Visualize variables

2 / 2 pts

✓ **+ 2 pts** Correct

+ 1 pt Visualizations aren't interpretable / missing some plots

+ 1 pt No description of the type of variables

+ 0 pts No visualizations

Question 4

Pre-process **4** / 4 pts

4.1 **Pre-process** **2** / 2 pts

✓ **+ 2 pts** Correct

+ 1 pt Partial Credit

+ 0 pts Incorrect/Missing

4.2 **Datasets** **2** / 2 pts

✓ **+ 2 pts** Dataset(s) for both problems

+ 1 pt Dataset(s) for one problem

+ 0 pts Incorrect or missing.

Q1 Data

3 Points

More details can be found on Canvas:

Q1.1 Dataset

1 Point

Briefly describe your dataset:

We picked the MushRoom Dataset. It contains 8124 instances of gilled mushrooms from the Agaricus and Lepiota families. Each instance has 22 categorical features, mainly physical characteristics such as cap shape, surface, color, bruises, odor, and gill properties. The target variable is the "poisonous" column of the dataset that is the mushroom being either "poisonous" or "edible". The dataset contains some missing values, and it is recommended for classification tasks.

OR

Upload a file that includes your definitions:

No files uploaded

Q1.2 Variables
2 Points

Briefly describe the variables in your dataset:

Class labels:
- poisonous: whether the mushroom is edible=e, or poisonous=p

Feature variables:
- cap-shape: General shape of the mushroom cap, like convex or flat.
- cap-surface: Texture of the cap surface, like scaly or smooth.
- cap-color: Color of the mushroom cap.
- bruises: Whether the mushroom shows visible bruising.
- odor: The scent of the mushroom
- gill-attachment: Describes how the gills are attached to the stalk.
- gill-spacing: Rhe density of spacing between the gills.
- gill-size: Width classification of the gills (broad or narrow).
- gill-color: Color of the gills under the cap.
- stalk-shape: Shape of the stalk, such as tapering or enlarging.
- stalk-root: Type or presence of root-like structures at the stalk base; may be missing.
- stalk-surface-above-ring: Texture of the stalk surface above the ring.
- stalk-surface-below-ring: Texture of the stalk surface below the ring.
- stalk-color-above-ring: Color of the stalk above the ring.
- stalk-color-below-ring: Color of the stalk below the ring.
- veil-type: Type of veil covering the mushroom; typically uniform in this dataset.
- veil-color: Color of the veil that covers the mushroom before it opens.
- ring-number: Number of rings present on the stalk.
- ring-type: The shape or structure of the ring on the stalk.
- spore-print-color: Color of the spores left behind when the mushroom cap is placed gill-down.
- population: Frequency at which the mushroom is found in the wild.
- habitat: Natural environment where the mushroom is typically found (e.g., woods, paths, grass).

OR

Upload a file that includes your definitions:

No files uploaded

Q2 Define a real-world motivation

4 Points

A classification model could be fit to this dataset.

Pick one or two response variables:

The class label 'poisonous'.

Who would want this model and why? What kind of "action" would they be able to take in order to increase profit, prevent future losses, or gain some other benefit?

It would be useful to foragers, nutritionists, and mushroom farming companies. Foragers and mushroom eaters could use this model to avoid the consumption of poisonous mushrooms during adventures. Mushroom agricultural companies can use this model to develop quality control pipelines to prevent poisoning the consumers with their sold mushroom products.

Give a reason for training multiple models for the same problem:

Training multiple models allows us to evaluate different aspects of model performance, such as predictive accuracy, interpretability, and reliability. Different model types can be good at different areas; for example, a random forest may be more accurate, but a decision tree gives greater interpretability.

OR

Upload a file that includes your answers:

No files uploaded

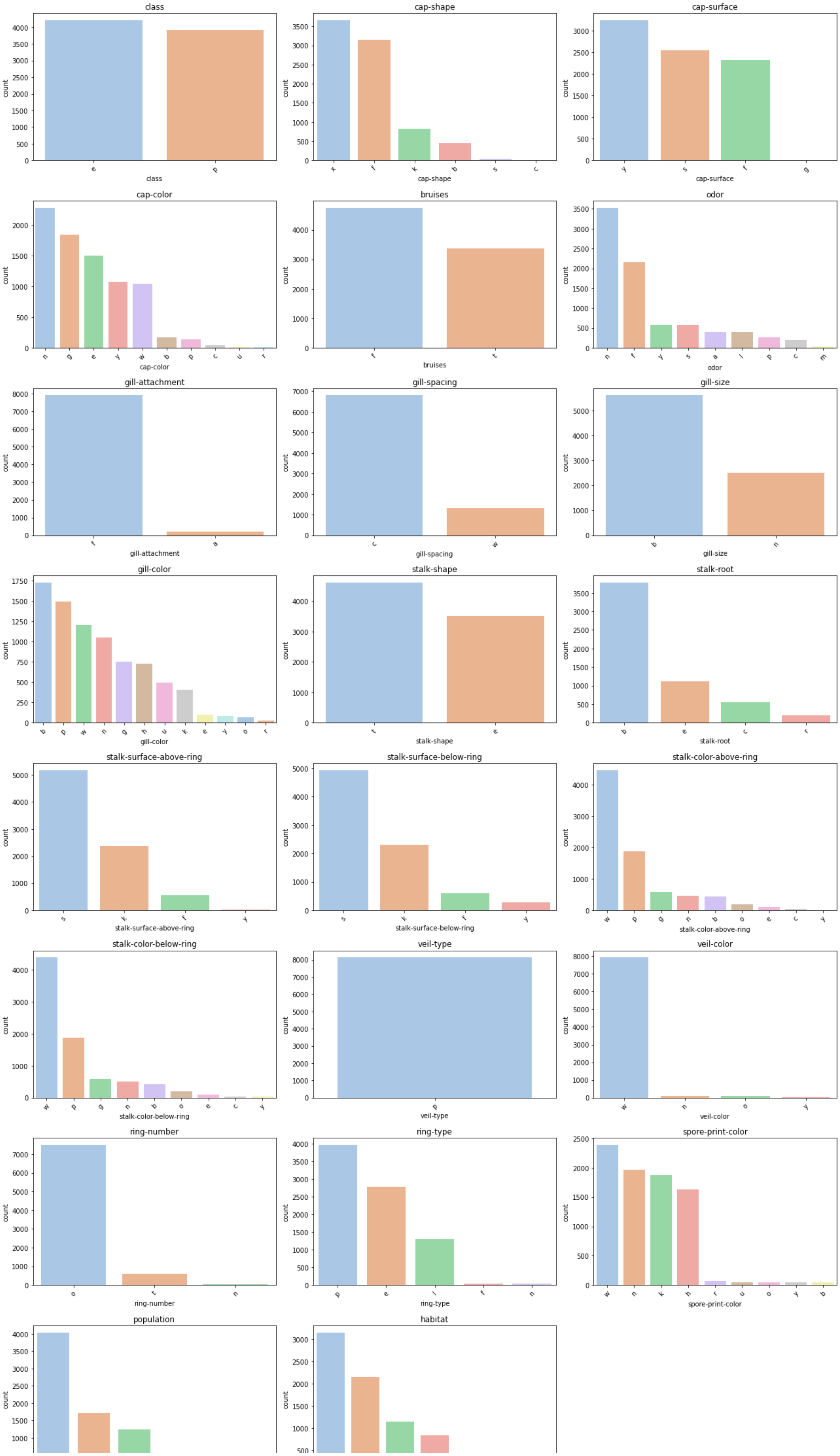
Q3 Visualize variables

2 Points

Create a visualization of the variables and upload them here:

download.png

Download


class
count
0
500
1000
1500
2000
2500
3000
3500
4000
e
p
class
cap-shape
count
0
500
1000
1500
2000
2500
3000
3500
x
f
k
b
s
c
cap-shape
cap-surface
count
0
500
1000
1500
2000
2500
3000
y
s
f
g
cap-surface
cap-color
count
0
500
1000
1500
2000
n
g
e
y
w
b
p
c
u
r
cap-color
bruises
count
0
1000
2000
3000
4000
f
t
bruises
odor
count
0
500
1000
1500
2000
2500
3000
3500
n
f
y
s
a
l
p
c
m
odor
gill-attachment
count
0
1000
2000
3000
4000
5000
6000
7000
8000
f
a
gill-attachment
gill-spacing
count
0
1000
2000
3000
4000
5000
6000
7000
c
w
gill-spacing
gill-size
count
0
1000
2000
3000
4000
5000
b
n
gill-size
gill-color
count
0
250
500
750
1000
1250
1500
1750
b
p
w
n
g
h
u
k
e
y
o
r
gill-color
stalk-shape
count
0
1000
2000
3000
4000
t
e
stalk-shape
stalk-root
count
0
500
1000
1500
2000
2500
3000
3500
b
e
c
r
stalk-root
stalk-surface-above-ring
count
0
1000
2000
3000
4000
5000
s
k
f
y
stalk-surface-above-ring
stalk-surface-below-ring
count
0
1000
2000
3000
4000
5000
s
k
f
y
stalk-surface-below-ring
stalk-color-above-ring
count
0
1000
2000
3000
4000
w
p
g
n
b
o
e
c
y
stalk-color-above-ring
stalk-color-below-ring
count
0
1000
2000
3000
4000
w
p
g
n
b
o
e
c
y
stalk-color-below-ring
veil-type
count
0
1000
2000
3000
4000
5000
6000
7000
8000
p
veil-type
veil-color
count
0
1000
2000
3000
4000
5000
6000
7000
8000
w
n
o
y
veil-color
ring-number
count
0
1000
2000
3000
4000
5000
6000
7000
o
t
n
ring-number
ring-type
count
0
500
1000
1500
2000
2500
3000
3500
4000
p
e
l
f
n
ring-type
spore-print-color
count
0
500
1000
1500
2000
2500
w
n
k
h
r
u
o
y
b
spore-print-color
population
count
1000
1500
2000
2500
3000
3500
4000
habitat
count
500
1000
1500
2000
2500
3000

Are any of them categorical? normally distributed? Which ones?

All variables in this dataset are categorical. Since there are no numeric variables, none are normally distributed.

However, visualizing the frequency distribution of each of these categorical variables helps us identify the dominant categories and potential data imbalance.

Q4 Pre-process
4 Points

Process the dataset to create two classification problems.

Q4.1 Pre-process
2 Points

Briefly explain how you defined two distinct classification problems for the same dataset:

- Predict whether the mushroom is edible or poisonous using the class label based on the mushroom's physical characteristics. This classification helps us avoid high-risk mushrooms.
- Predict the oder of the mushroom based on the other features. The potential fact that an olfactory feature could be predicted from physical features is interesting. It is practically useful for species identification and could be correlated with edibility.

OR

Upload a file that includes your answers:

🗎 No files uploaded

Q4.2 Datasets
2 Points

Upload your data files here:

▾ **mushroom.zip** Download

1 Large file hidden. You can download it using the button above.