Mushroom Edibility Classification

YANFEIYUN WU (KATHY)

DATA SCIENCE INSTITUTE AT BROWN UNIVERSITY

OCT 19TH, 2023

HTTPS://GITHUB.COM/KATHYWU1201/MUSHROOM_EDIBILITY_CLASSIFICATION_ML_PROJECT

I want to know…

- If there exists a mushroom that has never seen before, will it be edible or it is poisonous?
- As the nature is evolving, new type of fungi come out, but how do we know if we are safe to eat it?
- Help them to identify if the mushroom is safe to consume.
- Reduce the risk of death and intoxication due to poisonous mushrooms.

Data Descriptions

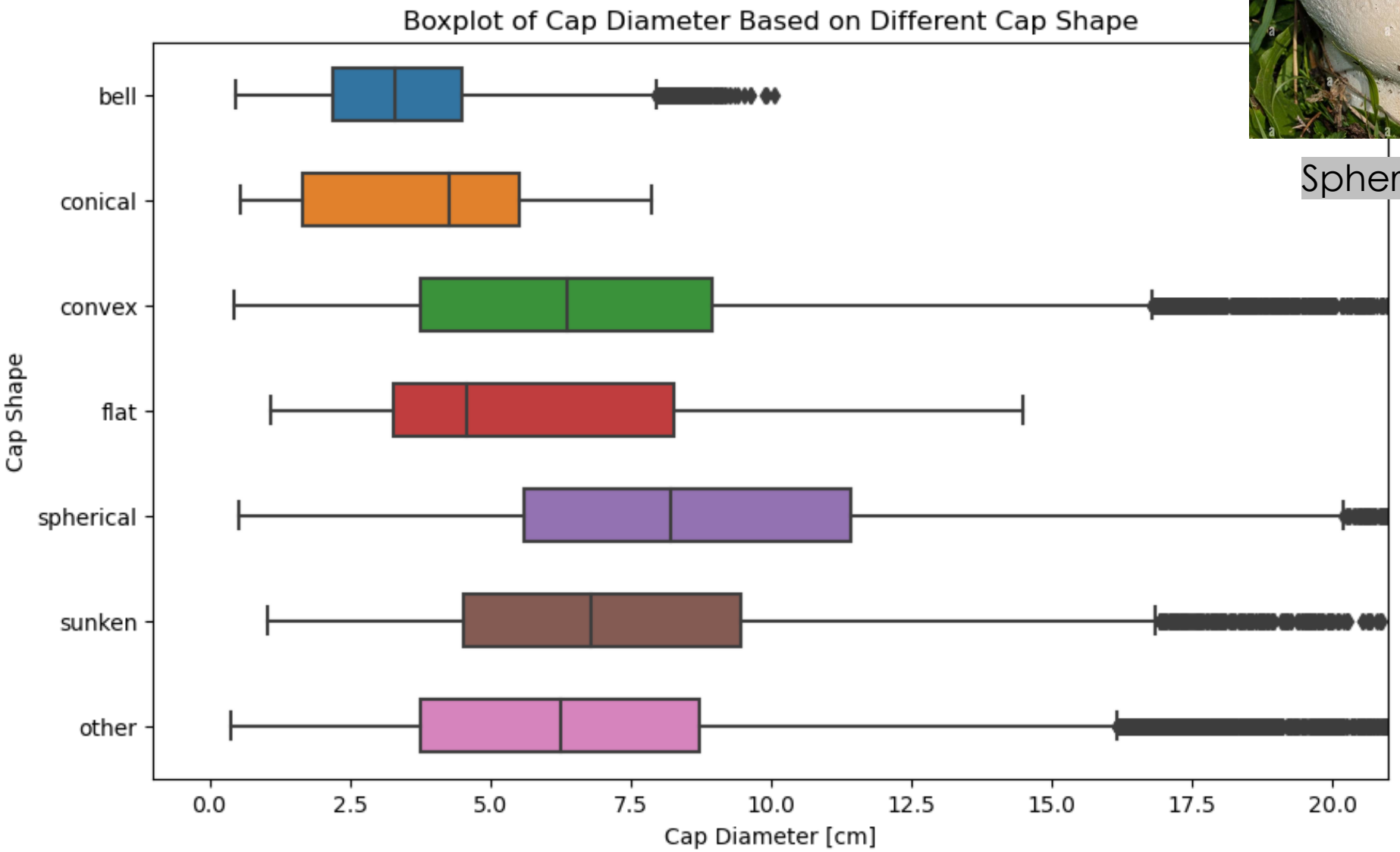
- Classification problem identifying if a new fungi is poisonous or not.
- ‘class’ feature as Target Variable.
- Data from Kaggle, collected from Patrick Hardin’s Mushrooms & Toadstools, and inspired by Jeff Schlimmer’s Mushroom Data Set.
- Original Data: 61069 rows, 21 columns.

https://www.kaggle.com/datasets/devzohaib/mushroom-edibility-classification/data

Boxplot of Cap Diameter

Bell shape

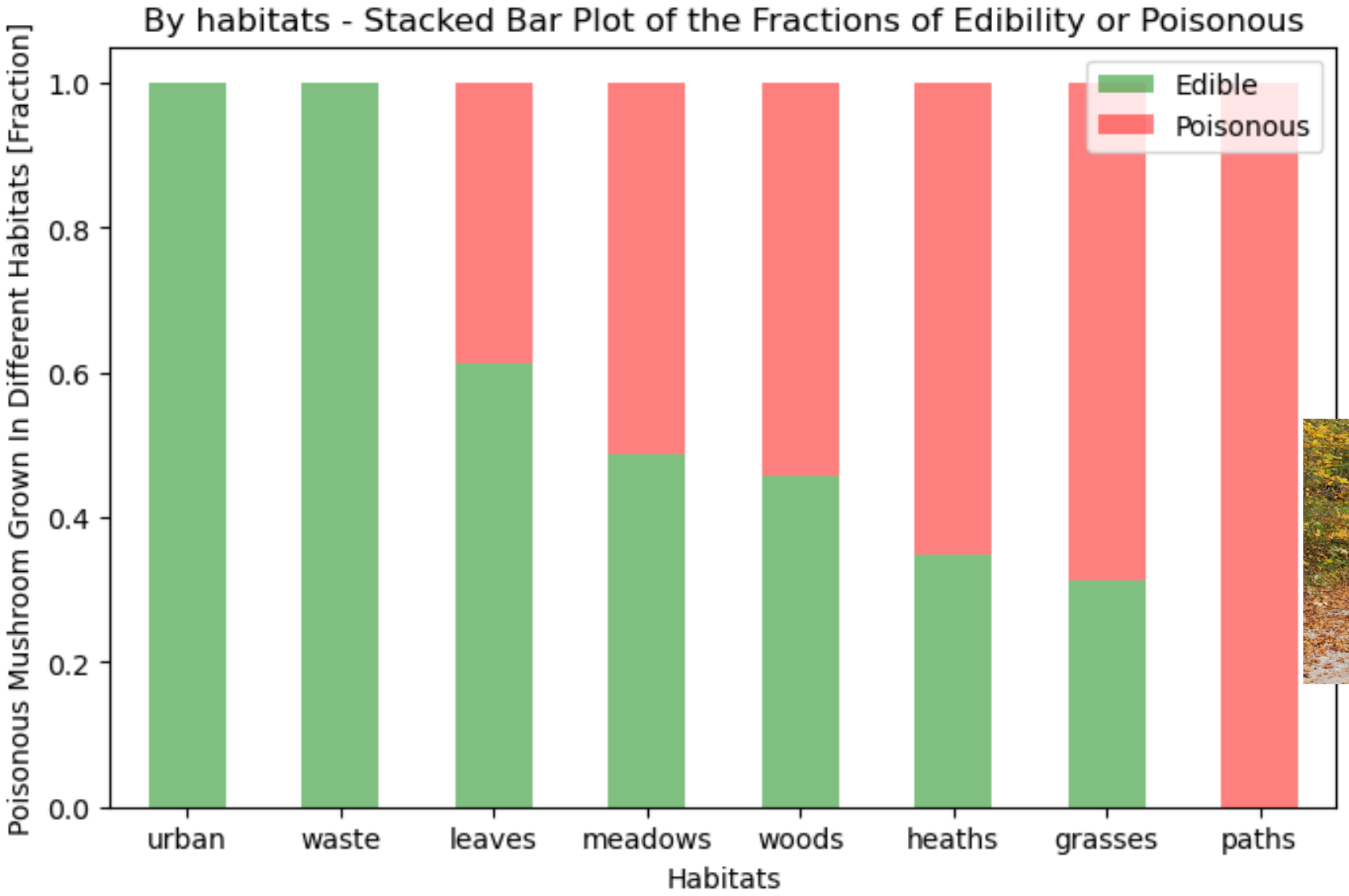
Spherical shape



@Dreamstime, https://thumbs.dreamstime.com/b/bell-shaped-mushrooms-12113748.jpg
@Alamy Stock Photo, https://c8.alamy.com/comp/2M3R75X/giant-puffball-giant-stag-fungus-spherical-whitish-fruiting-body-in-green-meadow-2M3R75X.jpg

By Habitats - Stacked Bar Plot

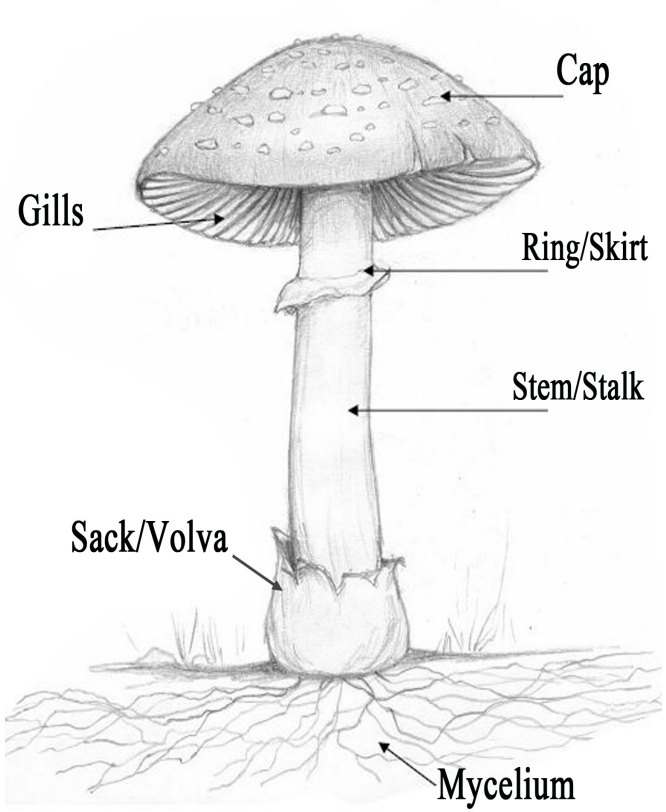
Urban habitat

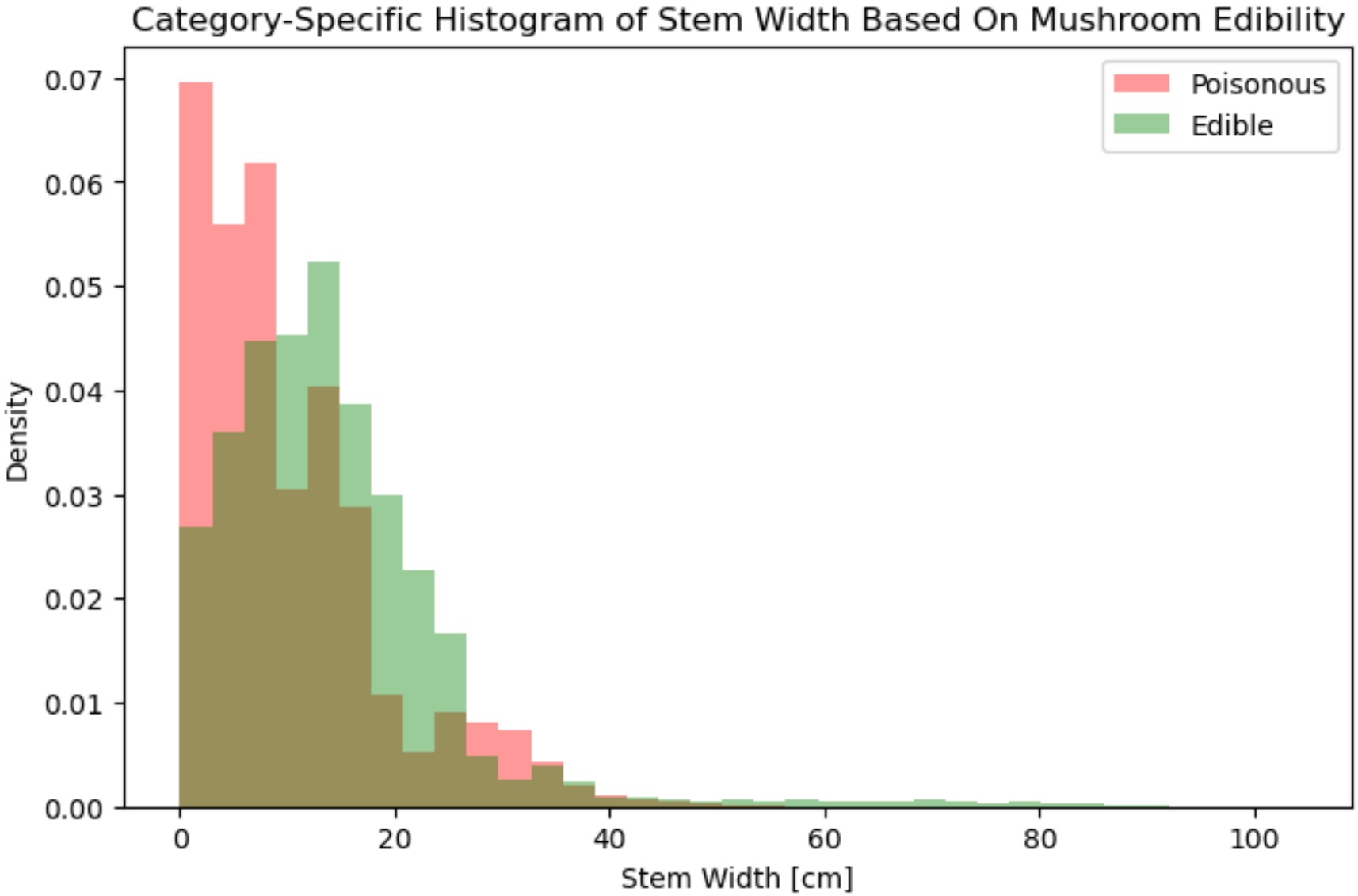


Path habitat

@Botanical Rambles, https://www.wnps.org/content/images/blog_articles/312/b2ap3_large_stalkedpuffball_brick.png
@Alamy, https://c8.alamy.com/comp/MB1YAD/path-running-through-mixed-forest-habitat-tara-national-park-serbia-october-MB1YAD.jpg

Category-Specific Histogram of Stem Width





@fungitownpod, https://fungitownpod.home.blog/2018/03/14/episode-02-how-to-id-mushrooms/

Splitting

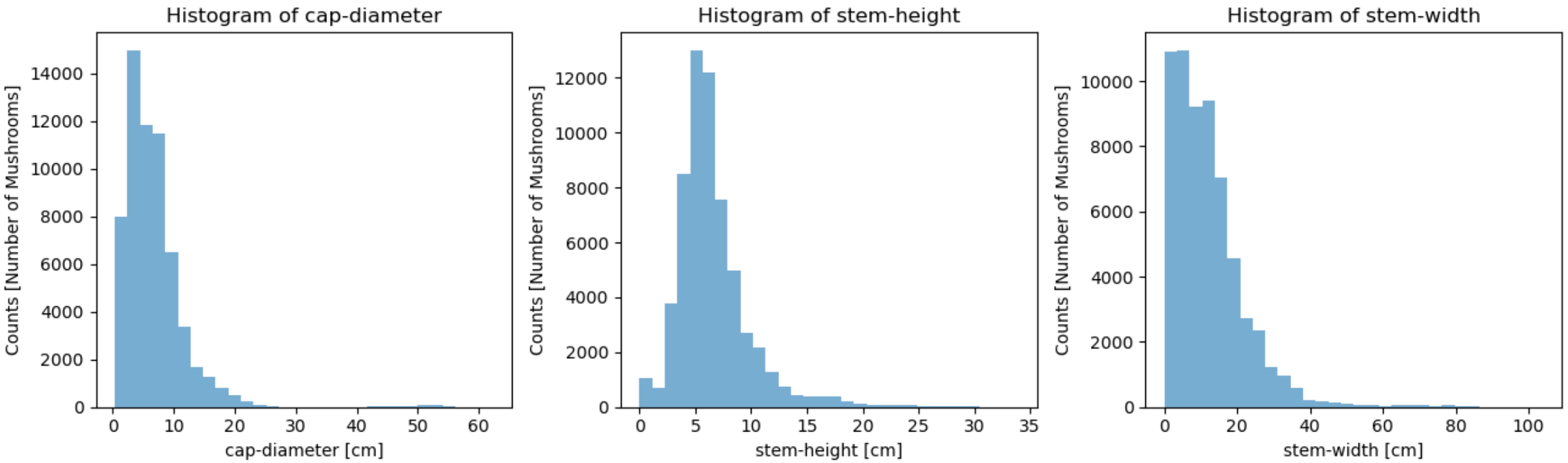
- The ratio of target variable (classification) is about 4:6 for edible and poisonous respectively.
- Use general **train_test_split** to set 20% of the dataset as test set and the rest 80% as other for **Kfold Cross Validation**.
- The shape of each other and test set are the following:
  X_other: (48855, 20), **X_test:** (12214, 20)
  y_other: (48855,) , **y_test:** (12214,)

Preprocessing

- 2 preprocessors are used:
  (i) Unorder categorical data: **one-hot encoder**
  (ii) Continuous features: **Standard Scalar**
- Most of the features are categorical data and unordered: e.g. color, shape, surface, etc.
- Two binary features - **does-bruise-or-bleed** and **has-ring** – will not perform any encoders.

Standard Scalar

- **cap-diameter**, **stem-height**, **stem-width**
- The features are heavy tailed, thus choose standard scalar



Missing Values

- Among **20 features** (exclude the target variable), **9 features** contains missing values.
- All the missing values are categorical features.
- **stem-root**, **veil-type**, **veil-color**, **spore-print-color** have more than 80% of missing values

```
fraction of missing values in features:
cap-surface          0.231214
gill-attachment      0.161850
gill-spacing         0.410405
stem-root            0.843931
stem-surface         0.624277
veil-type            0.947977
veil-color           0.878613
ring-type            0.040462
spore-print-color    0.895954
```

After Splitting and Preprocessing…

- Number of features **before** transformation: 20
  Number of features **after** transformation: 126
- Most of the features are categorical data, thus the number of features after preprocessing **increases**
- **No features** contain missing values after transformation


Thanks for listening!