Data Science: Capstone CYO Project - Mushroom

Elvin Tam

1 May 2021

CREDIT: GETTY IMAGES

Introduction

In this report, our goal is to predict the edibility (class: edible / poisonous) of mushroom basing on attribution information. Data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). The reason of selecting this dataset is that this problem is related to classification which is a large part of application in data science. And, it is also a complement to project – MovieLens that we can cover each part of what we have learnt from the course.

The mushroom dataset has already been well formatted from the source already. Data cleaning is only applied by removing 2 attributes prior to splitting the data to training set and test set. 10 algorithms are applied and an ensemble model combining the prior 10 different algorithms to see if it can provide improvement to our predictions.

1. glm
2. lda
3. Naïve Bayes
4. svmLinear
5. classification
6. knn
7. gamLoess
8. multinom
9. rf
10. adaboost
11. ensemble

1. Data Cleaning

Mushroom data set contains 23 columns of 1 class and 22 attributes related to cap, bruises, odor, gill, stalk, veil, ring, spore color, population and habitat of 8,124 observations.

```
## 'data.frame':    8124 obs. of  23 variables:
##  $ class                    : Factor w/ 2 levels "e","p": 2 1 1 2 1 1 1 1 2 1 ...
##  $ cap_shape                : Factor w/ 6 levels "b","c","f","k",..: 6 6 1 6 6 6 1 1 6 1 ...
##  $ cap_surface              : Factor w/ 4 levels "f","g","s","y": 3 3 3 4 3 4 3 4 4 3 ...
##  $ cap_color                : Factor w/ 10 levels "b","c","e","g",..: 5 10 9 9 4 10 9 9 9 10 ...
##  $ bruises                  : Factor w/ 2 levels "f","t": 2 2 2 2 1 2 2 2 2 2 ...
##  $ odor                     : Factor w/ 9 levels "a","c","f","l",..: 7 1 4 7 6 1 1 4 7 1 ...
##  $ gill_attachment          : Factor w/ 2 levels "a","f": 2 2 2 2 2 2 2 2 2 2 ...
##  $ gill_spacing             : Factor w/ 2 levels "c","w": 1 1 1 1 2 1 1 1 1 1 ...
##  $ gill_size                : Factor w/ 2 levels "b","n": 2 1 1 2 1 1 1 1 2 1 ...
##  $ gill_color               : Factor w/ 12 levels "b","e","g","h",..: 5 5 6 6 5 6 3 6 8 3 ...
##  $ stalk_shape              : Factor w/ 2 levels "e","t": 1 1 1 1 2 1 1 1 1 1 ...
##  $ stalk_root               : Factor w/ 5 levels "?","b","c","e",..: 4 3 3 4 4 3 3 3 4 3 ...
##  $ stalk_surface_above_ring: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ stalk_surface_below_ring: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ stalk_color_above_ring  : Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ stalk_color_below_ring  : Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ veil_type                : Factor w/ 1 level "p": 1 1 1 1 1 1 1 1 1 1 ...
##  $ veil_color               : Factor w/ 4 levels "n","o","w","y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ ring_number              : Factor w/ 3 levels "n","o","t": 2 2 2 2 2 2 2 2 2 2 ...
##  $ ring_type                : Factor w/ 5 levels "e","f","l","n",..: 5 5 5 5 1 5 5 5 5 5 ...
##  $ spore_print_color        : Factor w/ 9 levels "b","h","k","n",..: 3 4 4 3 4 3 3 4 3 3 ...
##  $ population               : Factor w/ 6 levels "a","c","n","s",..: 4 3 3 4 1 3 3 4 5 4 ...
##  $ habitat                  : Factor w/ 7 levels "d","g","l","m",..: 6 2 4 6 2 2 4 4 2 4 ...
```

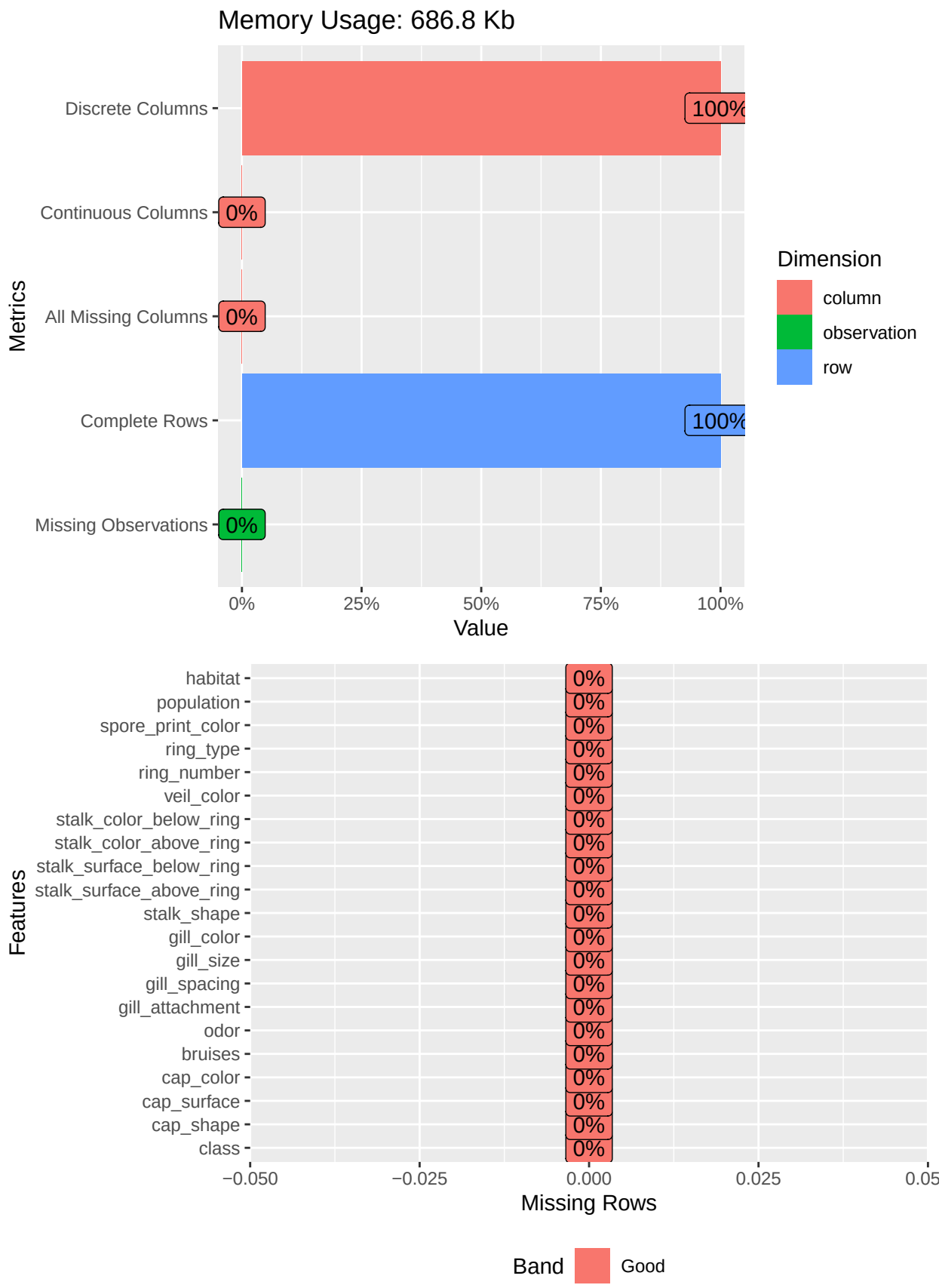
```
##  class    cap_shape cap_surface   cap_color    bruises       odor
##  e:4208   b: 452    f:2320      n      :2284   f:4748   n      :3528
##  p:3916   c:   4    g:   4      g      :1840   t:3376   f      :2160
##           f:3152    s:2556      e      :1500            s      : 576
##           k: 828    y:3244      y      :1072            y      : 576
##           s:  32                w      :1040            a      : 400
##           x:3656                b      : 168            l      : 400
##                                 (Other): 220            (Other): 484
##  gill_attachment gill_spacing gill_size   gill_color   stalk_shape stalk_root
##  a: 210          c:6812       b:5612    b      :1728   e:3516      ?:2480
##  f:7914          w:1312       n:2512    p      :1492   t:4608      b:3776
##                                         w      :1202               c: 556
##                                         n      :1048               e:1120
##                                         g      : 752               r: 192
##                                         h      : 732
##                                         (Other):1170
##  stalk_surface_above_ring stalk_surface_below_ring stalk_color_above_ring
##  f: 552                   f: 600                   w      :4464
##  k:2372                   k:2304                   p      :1872
##  s:5176                   s:4936                   g      : 576
##  y:  24                   y: 284                   n      : 448
##                                                    b      : 432
##                                                    o      : 192
##                                                    (Other): 140
```

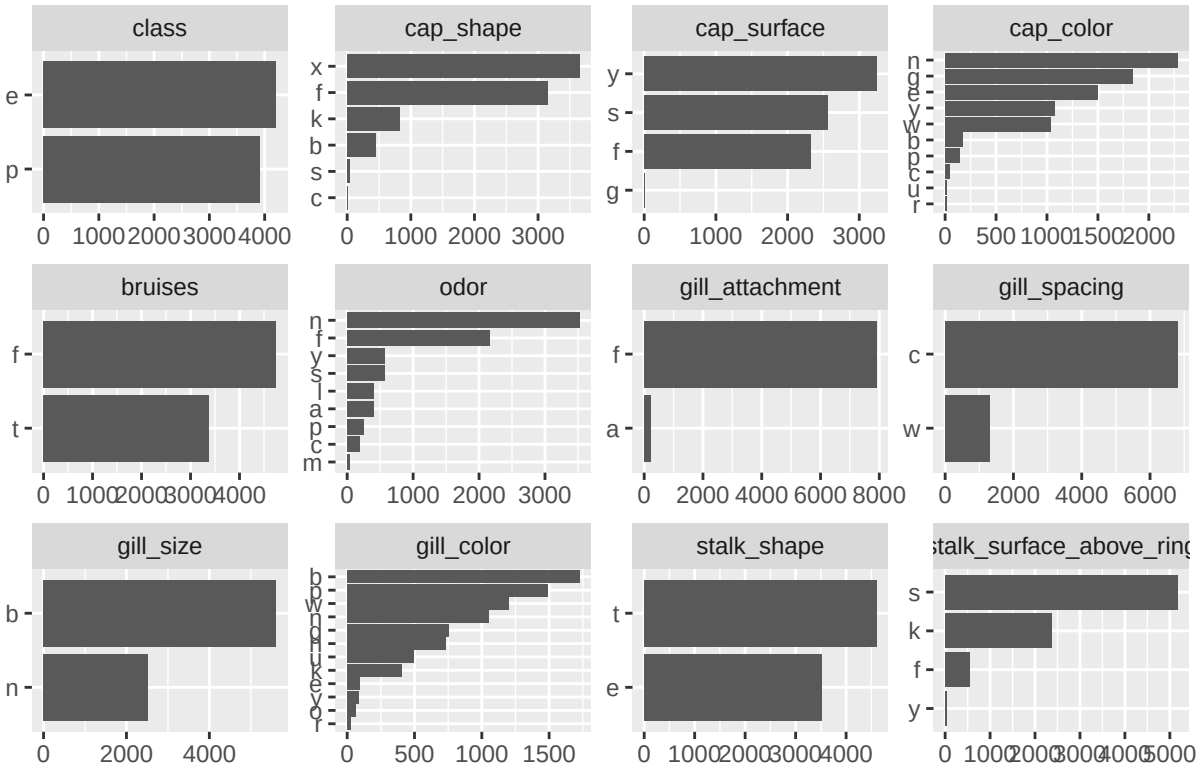
```
##  stalk_color_below_ring veil_type veil_color ring_number ring_type
##  w      :4384           p:8124    n:  96     n:  36      e:2776
##  p      :1872                     o:  96     o:7488      f:  48
##  g      : 576                     w:7924     t: 600      l:1296
##  n      : 512                     y:   8                 n:  36
##  b      : 432                                            p:3968
##  o      : 192
##  (Other): 156
##  spore_print_color population habitat
##  w      :2388      a: 384     d:3148
##  n      :1968      c: 340     g:2148
##  k      :1872      n: 400     l: 832
##  h      :1632      s:1248     m: 292
##  r      :  72      v:4040     p:1144
##  b      :  48      y:1712     u: 368
##  (Other): 144                 w: 192
```

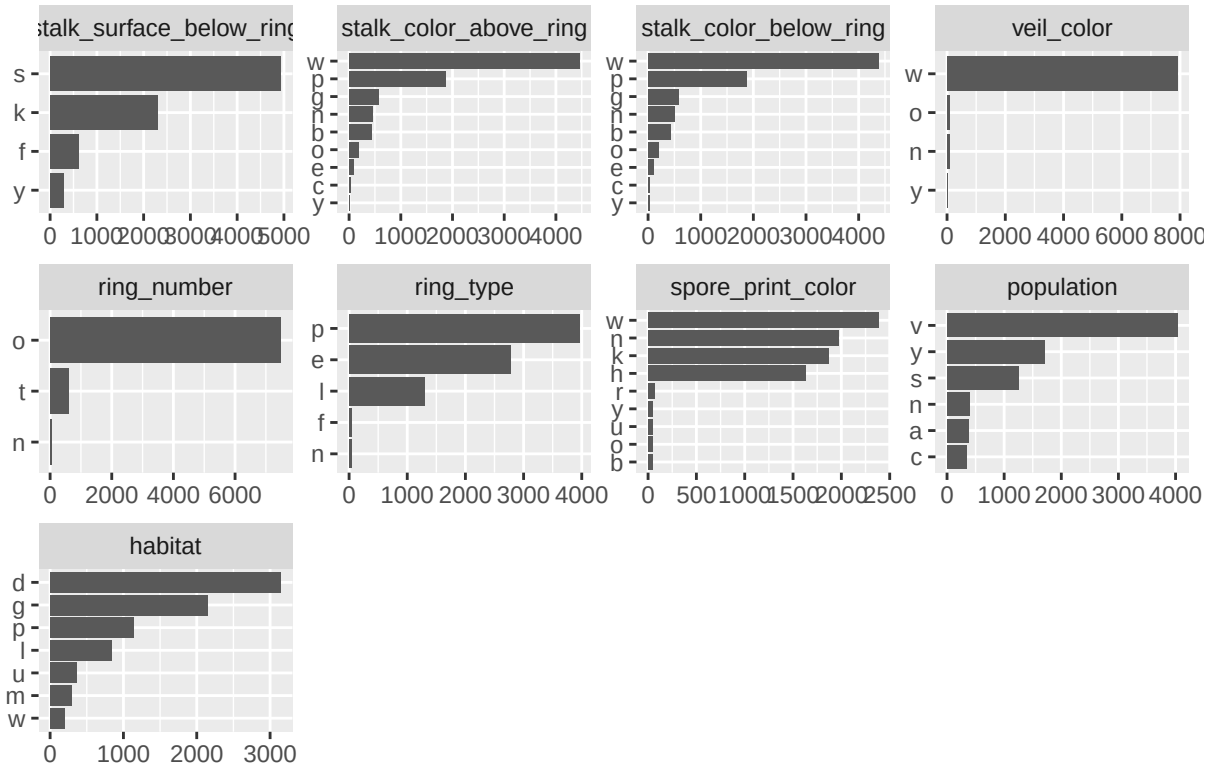
According to description from the source, there is data missing in the attribute of stalk root. The missing data point is marked "?" from the source already. Veil type is reported 1 level only. As a result, both stalk root and veil type are removed prior to we start data exploration & modeling.

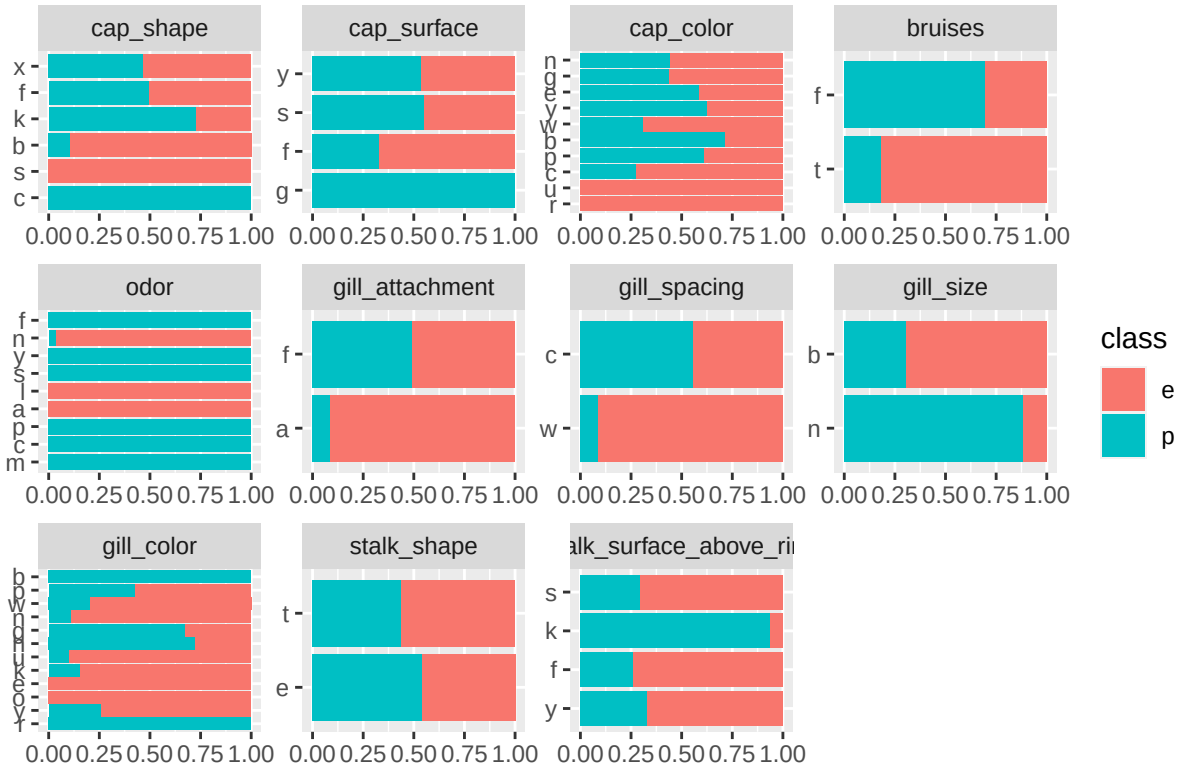
```
mushroom <- mushroom %>% select(-veil_type, -stalk_root)
```

2. Data Exploration

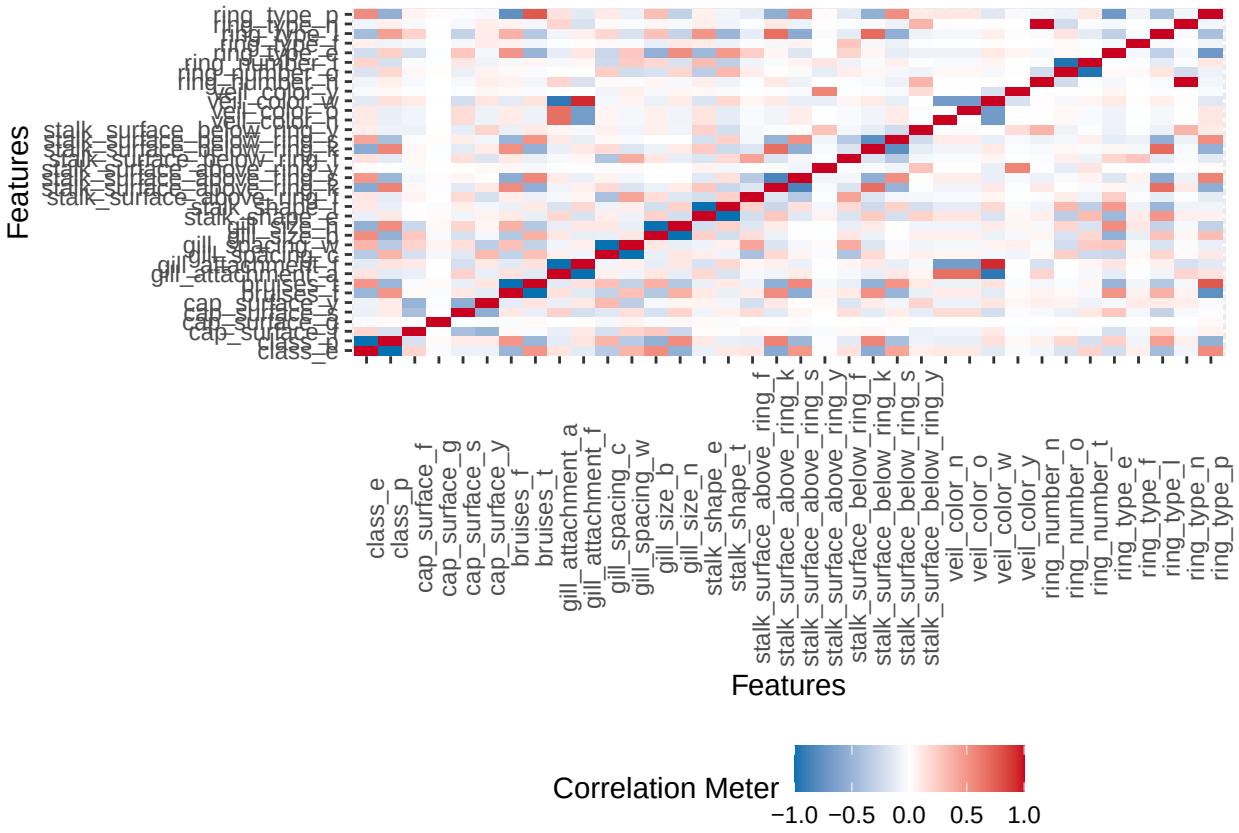



class
e
p
0 1000 2000 3000 4000
cap_shape
x
f
k
b
s
c
0 1000 2000 3000
cap_surface
y
s
f
g
0 1000 2000 3000
cap_color
n
g
e
y
w
b
p
c
u
r
0 500 1000 1500 2000
bruises
f
t
0 1000 2000 3000 4000
odor
n
f
y
s
l
a
p
c
m
0 1000 2000 3000
gill_attachment
f
a
0 2000 4000 6000 8000
gill_spacing
c
w
0 2000 4000 6000
gill_size
b
n
0 2000 4000
gill_color
b
p
w
n
g
h
u
k
e
y
o
r
0 500 1000 1500
stalk_shape
t
e
0 1000 2000 3000 4000
talk_surface_above_ring
s
k
f
y
0 1000 2000 3000 4000 5000
Frequency


stalk_surface_below_ring
stalk_color_above_ring
stalk_color_below_ring
veil_color
ring_number
ring_type
spore_print_color
population
habitat
Frequency


cap_shape
x
f
k
b
s
c
cap_surface
y
s
f
g
cap_color
n
g
e
y
w
b
p
c
u
r
bruises
f
t
odor
f
n
y
s
l
a
p
c
m
gill_attachment
f
a
gill_spacing
c
w
gill_size
b
n
gill_color
b
p
w
n
g
h
u
k
e
o
y
r
stalk_shape
t
e
alk_surface_above_ri
s
k
f
y
0.00 0.25 0.50 0.75 1.00
class
e
p


```
## 9 features with more than 5 categories ignored!
## cap_shape: 6 categories
## cap_color: 10 categories
## odor: 9 categories
## gill_color: 12 categories
## stalk_color_above_ring: 9 categories
## stalk_color_below_ring: 9 categories
## spore_print_color: 9 categories
## population: 6 categories
## habitat: 7 categories
```



3. Modeling Approach

Result

a results section that presents the modeling results and discusses the model performance

Conclusion

a conclusion section that gives a brief summary of the report, its limitations and future work