# Project 8: Applied Theory & Practice I, Mushrooms

Joe Zeimen

## Can you generate summary statistics that help describe the data?

The data for the mushrooms is discretized; there are no continuous values, and very few ordinal values. So the most interesting summary statistics for this data are the quick statistics on amounts and ratios of different classes. There are 8124 data points in the data: 51.8 are poisonous, and 48.2 are edible. So, there is quite a bit of data, and an even distribution. Continuing to explore the data in WEKA's explorer there are some other interesting things we should be aware of. There are 31% missing attributes in the stalk-root column. All mushrooms have the same veil type. We can take that out of our considerations as the data is mined later.

One the odor attribute is very interesting because it very accurately separates the data into poisonous and edible. This could be useful for classification later.

## Can the edible and poisonous data objects be distilled into groups?

Using Weka explorer. We can cluster the data using k-means and a fairly good confusion matrix. It is not perfect, but one of the reasons why there are more errors might be because any mushrooms that had unknown edibility were classified as poisonous. So in reality the ones that are assigned to cluster 0 but have the class poisonous may indeed be edible. This clustering was done using Weka and Kmeans clustering with Euclidian distance metric.

```
    0    1  <-- assigned to cluster
 4184   24 | e
  846 3070 | p

Incorrectly clustered instances :  886.0      10.906  %
```

## Can a classification model be created that can predict whether a mushroom is edible or poisonous?

Yes, one-way to do this is a rule-based classifier. Again, after trying many different ways with Knime, I chose Weka to handle this for me. With 9 rules it can classify the mushrooms with 100% accuracy. Which, actually, is probably not desirable in this situation because we know that most likely there are some mushrooms that are classified as poisonous that are actually edible. Nevertheless a rule-based classifier is great for this nominal data. This was done by using the jrip rule based classifier in Weka. It is interesting to note that it easily picked up on the odor to quickly break

down the data as we observed before. It also appears to have pulled out poisonous and if it did not match any of the rules then it is edible.
JRIP:

```
(odor = f) => poisonus=p (2160.0/0.0)
(gill-size = n) and (gill-color = b) => poisonus=p (1152.0/0.0)
(gill-size = n) and (odor = p) => poisonous=p (256.0/0.0)
(odor = c) => poisonous=p (192.0/0.0)
(spore-print-color = r) => poisonus=p (72.0/0.0)
(stalk-surface-below-ring = y) and (stalk-surface-above-ring = k) =>
      poisonus=p (68.0/0.0)
(habitat = l) and (cap-surface = y) and (population = c) => poisonous=p
      (12.0/0.0)
(cap-surface = g) => poisonus=p (4.0/0.0)
 => poisonus=e (4208.0/0.0)
Confusion matrix:
    a    b   <-- classified as
 4208    0 |    a = e
    0 3916 |    b = p
```

## Do any anomalies exist in the dataset?

This one has caused the most trouble for me. There are many approaches that can be used to find anomalies, but in this context it is not obvious to me what meaning it might have. Since we are looking at whether a mushroom is poisonous or edible, errors in classification may be a good place to start looking for anomalies.

One place I feel may be an anomaly in the data is when looking at odor, very few mushrooms are poisonous when looking at mushrooms with no odor. When mushrooms with an odor almost always are poisonous except in exactly the case of the odor of anise and almond. This could be an anomaly in the data. Perhaps odor exactly defines if a mushroom is poisonous or not, and our data just doesn't show it.

## Can any association rules be generated from this dataset?

Not very elegantly, the data mining tools have no problem coming up with all kinds of rules. Trying to hone in on the simplest rules is more difficult. I think this is again due to the fact that odor plays a significant role in determining if it is poisonous or not. So if you have a frequent item set that has a specific odor in it, chances are its easy to decide that it is poisonous or not. Trying to narrow down the rules I get seem to just get closer to the rules generated for the classification.