Amanreet Bajwa

Project 8

1. Can you generate summary statistics that help describe the data?



Figure - Counts for the poisonous vs. edible

The first thing I did was put the data into Knime to generate the statistics which didn’t give me much in terms of results. One statistic I was able to display were the counts of each part of each attribute. For example, poisonous vs. edible mushrooms; there are more edible mushrooms (approximately 51.8% or 4208 mushrooms) in this dataset than there are poisonous (approximately 48.2% or 3916 mushrooms). For the dataset in general it is hard to generate summary statistics that describe the data, besides the counts. The reason it is hard to do this is because the data is nominal. But, the data can only be changed into binary. Since it either it does or it doesn’t have grooves in its cap-surface, for example. I decided to use a column transform called One2Many. This transform split up the data into many binary columns so that I could try running the summary statistics on it once again. So at this point the data has a separate column for each value of each attribute. The statistics generated with these new columns didn’t really help much either. They do show the number of mushrooms per feature, but I had that information before and since they’re binary the min and max don’t matter. Also, the standard deviation and variance aren’t all that useful. At this point, due to the features of the data, I think it’s safe to conclude that interesting summary statistics that describe this data cannot be generated.

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

For this question, my first thought was to try to cluster the data. My mind automatically went to K-means.

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

When I first began this problem I went through a couple different options of classification in my head. I thought about trying a decision tree, nearest neighbor, and a rule based classifier. I decided on a decision tree because the data can be discretized and, hopefully, predict whether a mushroom is edible or poisonous based on different attributes. At this point I loaded the data into Knime and added a decision tree learner node. I configured the node to use the first column as the class name, use gini impurity, and do it without pruning. The decision tree actually generated a pretty good classification. The topmost decision was based on column 5 which is the odor of the mushroom. Using just the discrete values of this column most of the data was split into poisonous or edible clearly. The only value it did not cleanly separate was an odor of None.

1. Do any anomalies exist in the dataset?

From the decision tree it seems like poisonous when a mushroom has no odor is an anomaly since there are 3,408 that are edible and only 120 that are poisonous.

Unknown have been made poisonous so they may be anomalies.

1. Can any association rules be generated from this dataset?

Any number of rules can be generated