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?

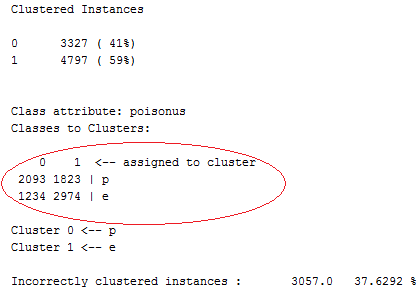


Figure - K-means clusters

For this question, my first thought was to try to cluster the data since clustering is essentially distilling data objects into groups. My mind automatically went to the K-means clustering algorithm. I already have the data loaded into Knime so I tried adding the k-means node to the workspace and connecting it to the csv reader node but that wasn’t working. Knime gave me a warning that says “No columns in include list! Produces one huge cluster.” I tried to edit the settings but since the attributes are nominal it wouldn’t let me add those to the list. I figured out how to add the Weka plugin for Knime so that I could use Weka nodes which do allow K-means to be used on nominal data. The SimpleKMeans node from Weka didn’t have this issue. I specified 2 clusters, hoping it would cluster poisonous in one and edible in another. This did not do so well. I got two clusters with 38% of the data in one and 62% in the other. I know from question 1 that the distribution of poisonous vs. edible is a little more even than that. I also tried this with the binary data and ended up with the same results. At this point I switched over to Weka and tried a SimpleKMeans with it ignoring the poisonous attribute. This actually distributed the data closer to the counts in problem 1 but, as you can see in Figure 2, it did not cluster the edible in one group and poisonous in the other, which was the idea of clustering it. I also spent a little bit of time trying to hierarchically cluster the data and use DBSCAN on it. These algorithms also resulted in no luck. At this point I do not think the data can be distilled into groups.

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 has already been made binary and, hopefully, this can predict whether a mushroom is edible or poisonous based on different attributes.

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