Association with Edible and Poisonous Mushrooms

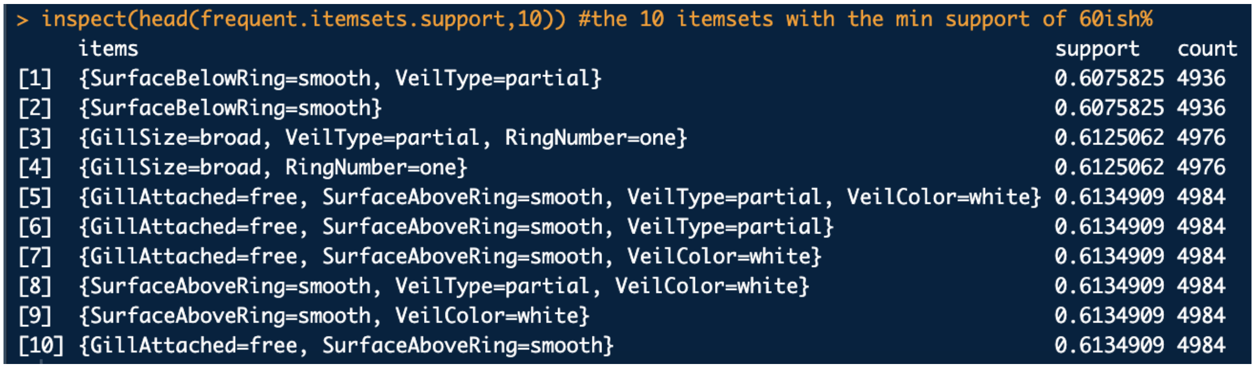
GOAL

To identify various association rules that provide us with characteristics and attributes that distinguish mushrooms as either edible or poisonous. Methods used will be K-means and hierarchical clustering.

ANALYSIS

The data mining process for association with the dataset “Mushrooms” begins with installing and loading the data and the two packages: “arules” and “arulesViz”. Then, moving on to the data exploration step, I summarized the dataset and found that there are 8124 mushrooms (or transactions) and 114 items (or columns). Using the inspect() function, I looked into the head and tail of the dataset and saw there are items such as odor, spore, habitat, and ring type. I also discovered that the itemsets in this data set don’t have the same number of items using the size() function. The size(head(Mushroom) was: 22 23 23 22 22 23. While the size(tail(Mushroom) was: 21 21 21 21 21 21. Since the sizes are not all the same, but fairly close to each other. This means that the items are less likely to be combined together and the chances of cooccurrences are lower than if the sizes were all the same.

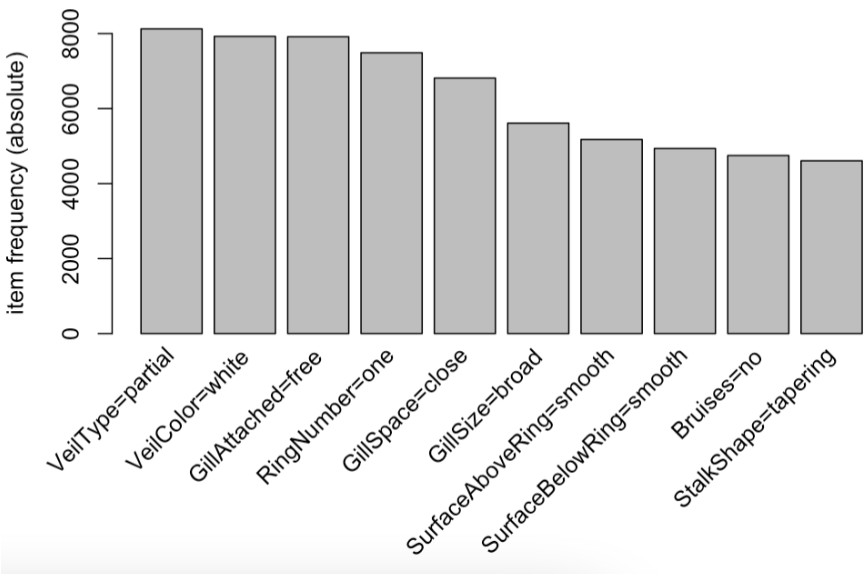
Afterward, I wanted to find the frequent itemsets, and based on the default support (ranging between 10-20%), there were 340,050 item sets. However, to get a better understanding of the itemsets and the relationship between items, I changed the minimum support to 60%. I used a larger minimum support of 60% since the length of each transaction is not very sparse and does not have high variability (being 22 23 23 22 21 21 etc). This also results in fewer item sets than compared to using a lower minimum support. So, I got 51 item sets and inspected the top 10 most frequent item sets by their support (this can be seen below).



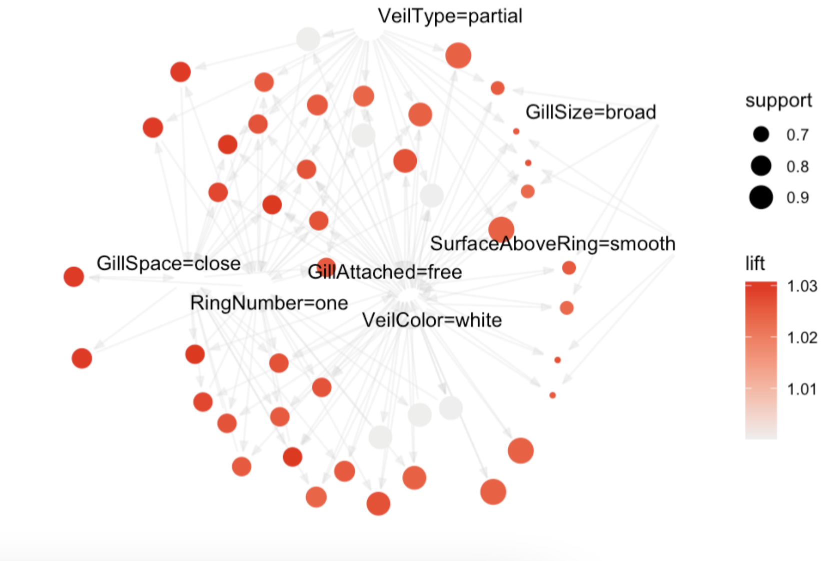
```
> inspect(head(frequent.itemsets.support,10)) #the 10 itemsets with the min support of 60ish%
     items                                                                        support   count
[1]  {SurfaceBelowRing=smooth, VeilType=partial}                                  0.6075825 4936
[2]  {SurfaceBelowRing=smooth}                                                    0.6075825 4936
[3]  {GillSize=broad, VeilType=partial, RingNumber=one}                           0.6125062 4976
[4]  {GillSize=broad, RingNumber=one}                                             0.6125062 4976
[5]  {GillAttached=free, SurfaceAboveRing=smooth, VeilType=partial, VeilColor=white} 0.6134909 4984
[6]  {GillAttached=free, SurfaceAboveRing=smooth, VeilType=partial}               0.6134909 4984
[7]  {GillAttached=free, SurfaceAboveRing=smooth, VeilColor=white}                0.6134909 4984
[8]  {SurfaceAboveRing=smooth, VeilType=partial, VeilColor=white}                 0.6134909 4984
[9]  {SurfaceAboveRing=smooth, VeilColor=white}                                   0.6134909 4984
[10] {GillAttached=free, SurfaceAboveRing=smooth}                                 0.6134909 4984
```

From this output, we can determine that SurfaceBelowRing=smooth, VeilType=partial, GillSize=broad, RingNumber=one, and the other highly frequent itemsets will be important to include when generating the rules. When the minimum support is increased, the number of itemsets that satisfy the support condition decreases which is why I got 51 itemsets. This can lead to higher confidence for the remaining 51 item sets. With a higher minimum support of 60%, I am hoping for stronger associations and higher confidence values.

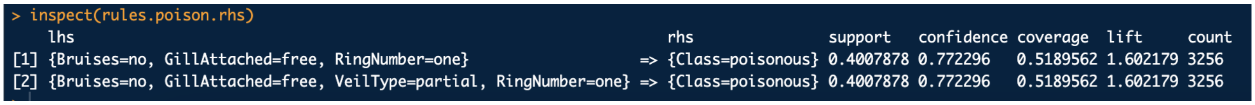
Using the itemFrequencyPlot(), I visualized the top 10 items to further filter and better understand which most frequent items to use for the association rules. The y-axis has the total count when each item occurs.



Moving on to generating the association rules! Using apriori() that applies the Apirori algorithm, I generated association rules with 60% minimum support and 75% minimum confidence since 75% is a good measure saying that 75% of the time there is an actual relationship between items. From this, 108 rules were returned, and I double-checked that a good handful of the rules had a lift greater than 1 because if the lift is equal to one or less than 1, the items are more likely to show up independently than together. I also sorted the item sets by support, and from this, I saw that with a support of 97%, confidence of 99%, and a lift of 1.024, a mushroom that has a free GillAttached will mostly have a VeilColor of white. It is important to note that when I changed the support to 50%, the lifts were still around 0.9-1.03. When I changed the confidence level to 85%, the lifts remained within this range. I argue that sticking with a support of 60% and a minimum confidence of 75% is still okay because it does give us rules in which the lift is greater than 1 even if it is close to being equal to one. As long as some lifts are greater than 1, it still shows a slightly higher possibility of the items happening together than independently. Therefore, I created a rules.lift <- subset(rules, lift>1) to focus on the rules that show the items will most likely happen together rather than apart. This narrowed it down to 46 rules and then I visualized these rules in the graph below. In this graph, we see that the items VeilType=partial and SurfaceAboveRing=smooth have a high lift and high support and should be taken into account for the rules.



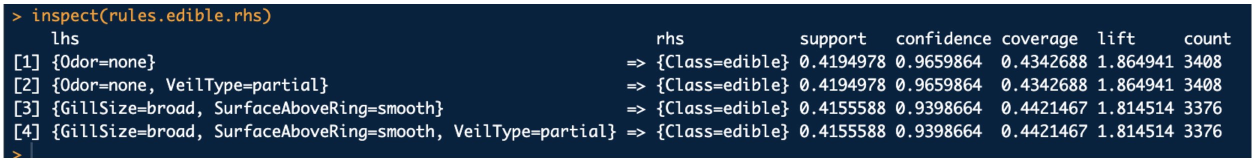
Finally, for the last part, I targeted specific LHS and RHS itemsets. I created 2 sets of rules for edible mushrooms and poisonous mushrooms. To begin with, I was wrong about my prediction of having a support of 60% and confidence of 75%. Zero rules were returned when I added more conditions to the rules. For the conditions of the rules for poisonous mushrooms, I wanted it to be a minimum of two items in the itemset and the RHS had to be “Class=poisonous”. Unfortunately, 60% support and 75% confidence did not return anything, so I had to make adjustments and change the support to 40%. It is important to keep the confidence high and ensure that the lift is greater than 1. After these changes, I got the following 2 rules:



```
> inspect(rules.poison.rhs)
    lhs                                                                  rhs               support   confidence coverage  lift     count
[1] {Bruises=no, GillAttached=free, RingNumber=one}                   => {Class=poisonous} 0.4007878 0.772296   0.5189562 1.602179 3256
[2] {Bruises=no, GillAttached=free, VeilType=partial, RingNumber=one} => {Class=poisonous} 0.4007878 0.772296   0.5189562 1.602179 3256
```

This shows that if a mushroom has no bruises, a free gill attached, one ring number, and/or a partial veil type, it will most likely be a poisonous mushroom. These adjustments also worked out because the lift increased significantly and differently showing that these items will happen together than independently.

For the edible mushrooms, I had the same issue with the 60% support and 75% confidence when I also had the other conditions of a minimum of two items in the itemset and the RHS as “Class=edible”. So, I made the same changes and with 40% support, 75% confidence, and a lift greater than 1, it returned the four following rules:



```
> inspect(rules.edible.rhs)
    lhs                                                            rhs            support   confidence coverage  lift     count
[1] {Odor=none}                                                 => {Class=edible} 0.4194978 0.9659864  0.4342688 1.864941 3408
[2] {Odor=none, VeilType=partial}                               => {Class=edible} 0.4194978 0.9659864  0.4342688 1.864941 3408
[3] {GillSize=broad, SurfaceAboveRing=smooth}                   => {Class=edible} 0.4155588 0.9398664  0.4421467 1.814514 3376
[4] {GillSize=broad, SurfaceAboveRing=smooth, VeilType=partial} => {Class=edible} 0.4155588 0.9398664  0.4421467 1.814514 3376
>
```

This shows that if the mushroom is odorless and/or has a partial veil type, and has a broad gill size and smooth surface above the ring and/or a partial veil type, the mushroom is mostly likely edible. Interestingly, the original rules showed in the graph that a partial veil type and the smooth surface above the ring were items that would play a role in the association rules.

TOOLS & RESOURCES

Language: R

Dataset information: “Mushrooms” dataset that is pre-installed in R Studio.