**Introduction**

*Contextual Background: History of Mushrooms*

The “Natural World” is a global term used to describe all components of the ecosystems on the planet. This could include bodies of water, populations of creatures, the inside of your stomach, or even the ground you walk on. Some might think the natural world is the antecedent to the manmade world, but this is simply not true. there is no real natural world, nor man-made world, Ecosystems are highly diverse, some arising completely independently of others due to the nature in which their components comingle to shape a living system. The only real reason there is a “Natural World”, is because there is now a manmade world. While it was not always this way, there

* Tie together mushrooms and humanity in a fancy but succinct way. Its simply true that

We know that the human species *(homo sapiens)* originally was part of an ecosystems with many other species of hominids present. The most common ancestor of them all is thought to have lived some 500,000 – 750,000 years ago, however there are several candidates that scientist are considering for the position. This is partly because hominids at large became incredibly spread and diversified. Many skeletal remains of confirmed homo sapiens species have been found across the globe, with some specimens dating back to 300,000 years in the infamous Jebel Irhoud, a site home to the oldest *homo sapiens* remains. (Handwerk, B. 2021) Although it is not always clear what set them apart from the other hominids, there is evidence to suggest *homo sapiens* had reached a significant number of ecosystems during prehistoric eras. With that came a long history of interactions with organisms and features of those systems. One such organism of incredible significance were mushrooms. Humans have been marveling at these organisms since the paleolithic age. They are referenced as “Food of The Gods”, by the Romans (Valverde, M. E., 2015), and revered highly by the Chinese since 600 B.C.E. (FoodPrint) Roman philosopher Galen, wrote in his notebook of the Chinese cultivating what where likely shitake mushrooms The Fungi kingdom represents a highly diverse population of organisms. A picture containing rock, stone

Description automatically generated

(Talk about the history of Humans and Mushrooms)

Given the difficulties present in natural species identification, it may not be possible to predict any mushrooms species without very discrete data. However, the inspirational dataset for this project enables the testing of various data mining techniques effectiveness on predicting mushroom edibility.

* Talk about both the species, show pictures
* Relate mushroom diversity

*Describing Mushrooms as “Equally Diverse”*

Within that are mushrooms which have equally diverse variation amongst themselves. Extensive vocabulary is needed to effectively describe the patterns of features that may occur across species. Additionally, Mushrooms will exhibit a wide variety of features in their own species that can make classification difficult to an untrained individual. Even with decently accurate feature identification, it is very difficult to accurately classify mushrooms due to overlap in morphology between species that are poisonous with the species that are edible. Often, correct identification requires an expert evaluation of the complex features that are observed with the naked eye. Yet still, most mushroom specimens found cannot be definitively classified due to the morphological overlaps present. Using microscopic tools with spore printing techniques is the most accurate method for identifying specimens. Granted, spore prints are only a glance and the physical attributes of the reproductive parts of a mushroom and can be similar across genetically distant species.

*Identifying Mushrooms*

Despite the conundrum of identification, the mycological field has been able to evolve strong identification methods of many mushrooms based on visual presentation. …

**Analysis**

*The Mushroom Dataset*

Investigating the identification of mushrooms through features can be done if enough examples of a wide variety of features are present. There is much work to be done in the field to digitize and apply machine learning algorithms. Despite that, the inspirational dataset was obtained from a [Kaggle](https://www.kaggle.com/datasets/uciml/mushroom-classification?resource=download), which uses hypothetical samples corresponding to 23 species of “gilled” mushrooms of the Agaricus and Lepiota Families drawn from the “Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as “definitely edible”, “definitely poisonous”, “unknown edibility” and “not recommended”. The data also covers 22 features of the mushrooms, coded as single letters. There are no missing or N/A categories in the 8125 rows of the dataset.

Taken from the link described above, as a description provided on Kaggle’s website of the data:

“Although this dataset was originally contributed to the UCI Machine Learning repository nearly 30 years ago, mushroom hunting (otherwise known as "shrooming") is enjoying new peaks in popularity. Learn which features spell certain death and which are most palatable in this dataset of mushroom characteristics. And how certain can your model be?”

“This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be'' for Poisonous Oak and Ivy.

* Time period: Donated to UCI ML 27 April 1987”

What is the problem for the dataset?

In this case we have a collection of 23 species of mushrooms of the AGaricus and Lepiota families. These species are classified as Edible or Poisonous ( and have the unsure but not recommended roped into the poisonous mushrooms category). There are no simple rules regarding the physiology of mushrooms that can confirm a poisonous or edible mushroom. Can we use clustering techniques to identify a poisonous or edible mushroom, or can we predict the species of mushroom based on classification?

*Pre-processing of Data*

While the mushrooms.csv file was generally clean, some generalized and model specific pre-processing was applied to enable the utilization of various classification and data mining techniques. For testing a models prediction (i.e. predict edibility) the data will require formatting that stores “chr” coded features as factors and removes all labels of “poisonous” or “edible” or otherwise “revealing” classifications. The data will also need to be separated into a training and test dataset for assessing error across models. Through that, model effectiveness is determined predicting a mushroom’s edibility. The applicable data mining techniques planned for testing are K-means algorithm clustering, Hierarchical clustering, Association Rule Mining, and Decision Tree Algorithm partitioning. Once models are trained and tested for each of these techniques, the final models are used to generate confusion matrices for various stages of the techniques. Confusion matrices permit examination of training accuracy, and final testing accuracy for models.

*Levels of the Features*

$class

[1] "e" "p"

(classes: edible=e, poisonous=p)

$cap.shape

[1] "b" "c" "f" "k" "s" "x"

bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s

$cap.surface

[1] "f" "g" "s" "y"

fibrous=f,grooves=g,scaly=y,smooth=s

$cap.color

[1] "b" "c" "e" "g" "n" "p" "r" "u" "w" "y"

brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w,yellow=y

$bruises

[1] "f" "t"

bruises=t,no=f

$odor

[1] "a" "c" "f" "l" "m" "n" "p" "s" "y"

almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s

$gill.attachment

[1] "a" "f"

attached=a,descending=d,free=f,notched=n

$gill.spacing

[1] "c" "w"

close=c,crowded=w,distant=d

$gill.size

[1] "b" "n"

broad=b,narrow=n

$gill.color

[1] "b" "e" "g" "h" "k" "n" "o" "p" "r" "u" "w" "y"

black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y

$stalk.shape

[1] "e" "t"

enlarging=e,tapering=t

$stalk.root

[1] "?" "b" "c" "e" "r"

bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?

$stalk.surface.above.ring

[1] "f" "k" "s" "y"

fibrous=f,scaly=y,silky=k,smooth=s

$stalk.surface.below.ring

[1] "f" "k" "s" "y"

fibrous=f,scaly=y,silky=k,smooth=s

$stalk.color.above.ring

[1] "b" "c" "e" "g" "n" "o" "p" "w" "y"

brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y

$stalk.color.below.ring

[1] "b" "c" "e" "g" "n" "o" "p" "w" "y"

brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y

$veil.type

[1] "p" “u”

partial=p,universal=u

$veil.color

[1] "n" "o" "w" "y"

brown=n,orange=o,white=w,yellow=y

$ring.number

[1] "n" "o" "t"

none=n,one=o,two=t

$ring.type

[1] "e" "f" "l" "n" "p" “c”, “s”, “z”

cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z

$spore.print.color

[1] "b" "h" "k" "n" "o" "r" "u" "w" "y"

black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y

$population

[1] "a" "c" "n" "s" "v" "y"

abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y

$habitat

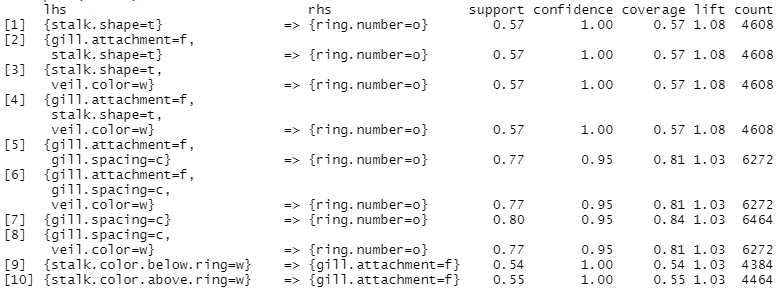
[1] "d" "g" "l" "m" "p" "u" "w"

grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

*Association Rule Mining*

Arules package in R is used to generate relationships between features within the processed mushrooms dataset. Additionally pre-processing for association rule mining included factorizing every categorical column representing a mushroom feature. A strong set of general rules was achieved through tuning the support, confidence, and maxlen parameters using the apriori algorithm for associative rule generation. A large set (98) of rules which encapsulates many of the features on the mushroom’s dataset, was generated to observe rules at large. The tuning for the apriori algorithm had parameters of support = 0.5, confidence = 0.9, maxlen = 5, and the top 10 sorted by decreasing lift are shown below.

*Large set of Association Rules (98)*



*Targeting Features with Association Rule Mining*

Through targeting class of “e” or “p” on both the left- and right-hand side of the rule, the following sets were generated. The tuning parameters were set to support = 0.30 , confidence = 0.90 , maxlen = 5 . The following rulesets show the top 10 ( if present ) for comparison.

***Target class=e on rhs***

***Text

Description automatically generated with low confidence***

Text

Description automatically generated

***Target class=e on lhs***

***A picture containing text

Description automatically generated***

Table

Description automatically generated

***Target class=p on rhs***

Graphical user interface

Description automatically generated with medium confidence Text

Description automatically generated

***Target class=p on lhs***

***A picture containing graphical user interface

Description automatically generated*** ***A picture containing table

Description automatically generated***

*Decision Tree and Random Forest*

Timeline

Description automatically generated Decision Trees were made after first performing testing and training split sequencing of the original data. Every 4th index went into the test dataset., leaving 75 % in the training data ( about 6000 instances) . Decision Tree makes use of the factorized dataframe that was pre-processed prior. The output of the call is shown to the right with the corresponding confusion matrix from prediciton displayed directly below.

Text

Description automatically generated with medium confidence

*Random Forest*

Given that random forest is an iteration of the Decision tree method, the same training and test data could be used. The ntree parmeter specifies the number of decision trees to use in the algorithm, with final number being set to ntree= 11 for 100% prediction accuracy.

Text

Description automatically generated

**Results**

*Results of Association Rule Mining Large Set (98 rules)*

Generating this large set of rules was important for visualizing associations of features, however it would be confusing if a researcher did not know there were two distinct classes of mushrooms present in the dataset. Notice with the larger dataset, sorting by decreasing lift reveals weak lift in a set of rules targeted at all features in the dataset. It is likely that while there are overlap in some features between classes of mushroom, we are unable to tell because of the strength with which one feature is associated with another. The maxlen = 5 parameter setting did nothing to limit the output in this instance, because there are usually only two features on the left-hand side of the rules. While this information does nothing to predict edible class, it is informative of the strengths between features in mushrooms. Below is a visualization of the set of top 10 rules association found when using parameters support = 0.30, confidence = 0.90, maxlen = 5 .

Chart

Description automatically generated

*Results of Targeting Features with Association Rule Mining*

In targeting features on the left- and right-hand side, it was most evident that mushrooms of a particular class were associated with certain features. Additionally, this analysis revealed that some features are globally associated with mushrooms of either class. Identical tuning parameters of support = 0.30, confidence = 0.90, maxlen = 5 were used when focusing on the left-hand side. While these generated disproportionate sizes of rules overall (class=’e’, 88 rules) vs (class=’p’, 18 rules), notice the support for the rules of strongest lift are all barely over support = 0.30. This was discovered during the tuning of the rules, and in looking at the features present in the top 10 sets allows us to understand what features are most important for mushrooms of either class, or which are important for neither. The visualizations for the top 10 rules of either class on targeted on the rhs are on the next page. These are features that associate with this class, and not the class’s association with features which is shown in the next section.

***Visualization of top 10 rules Targeting class=e on rhs***

**Chart

Description automatically generated**

***Visualization of top 10 rules Targeting class=p on rhs***

**A picture containing sky, green

Description automatically generated**

*Features Associated with Class ‘e’ or ‘p’*

The node diagrams above show features linked via strength through distance, notice first that gill.attachment=f, veil.color=w, and ring.attachment=o are universal in both sets of class predicting rules. It’s important to remember the data was from two sets of closely linked species, each having their own poisonous and non- members. Therefore, these features can be recognized as strongly correlative because of the general morphology of the species present. If this study was apply to mushrooms universally, there would be differ feature association results based on the species contained within the dataset analyzed. What is not the same between rhs association rules?

*Results of Decision Tree*

**Conclusions**

Overfitting of models will need to be addressed given the nature of feature overlap within the dataset. Conclusions will be drawn around the accuracy of models, and assessments will be made on whether or not data mining techniques can be used in this context.

**References**

FoodPrint. (2019, December 9). *Mushrooms*. FoodPrint. Retrieved September 5, 2022, from https://foodprint.org/real-food/mushrooms/

Handwerk, B. (2021, February 2). *An evolutionary timeline of homo sapiens*. Smithsonian.com. Retrieved September 5, 2022, from https://www.smithsonianmag.com/science-nature/essential-timeline-understanding-evolution-homo-sapiens-180976807/

Valverde, M. E., Hernández-Pérez, T., & Paredes-López, O. (2015). *Edible mushrooms: Improving human health and promoting Quality Life*. International journal of microbiology. Retrieved September 5, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4320875/#:~:text=Mushrooms%20have%20been%20consumed%20since,have%20been%20part%20of%20the

The **dataset**

<https://www.kaggle.com/datasets/uciml/mushroom-classification?resource=download>

Images used in powerpoint

*Agaricus Bisporus* (Slide 1)

<https://candide.com/img/39f5dd34-6b60-4444-b419-d939d3b00c6a/cropped/600x480>

(Inserted ) Slide 2

<https://i0.wp.com/curejoy.com/wp-content/uploads/2017/02/Poisonous-Mushrooms-You-Should-Stay-Away-From.jpg?fit=1200%2C627&ssl=1>

*Lepiota racchodes*  (Slide 2)

<http://urbanmushrooms.com/images/18.jpg>

Slide 3 Reddening Lepiota

<https://static3.bigstockphoto.com/0/0/2/large1500/200331292.jpg>

Slide 4 Stinking Dapperling

<https://www.wildfooduk.com/wp-content/uploads/2019/10/Cristata-9-1-720x540.jpg>

*Lepiota Scarabulla* (Slide 5)

<http://iucn.ekoo.se/assets/uploads/d3c7b-Lepiota-scaberula_SFWatershed_6-Dec-2014_NS.jpg>

*Agaricus Agruptabulbus* (Slide 6)

<https://www.fungikingdom.net/_Media/agaricus--abruptibulbus1066_med_hr.jpeg>

*Lepiota Clypeolaria* (Slide 8)

<https://upload.wikimedia.org/wikipedia/commons/5/5b/Lepiota_clypeolaria.JPG>

**Notes and Resources on Mushrooms**

Speciation in fungi

<https://academic.oup.com/bioscience/article/49/6/440/229441>