**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. Ecosystems are highly diverse, some arising completely independently of others due to the nature in which their components comingle to shape a living system. 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 their interaction with several The Fungi kingdom represents a highly diverse population of organisms.

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.

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. The Dataset was acquired from [Kaggle](https://www.kaggle.com/datasets/uciml/mushroom-classification?resource=download), 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.

In order to use the dataset to 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.This preparation will then permit us to use a training / testing approach when constructing models, to determine how well each of our various models were able to predict 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. 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.

**Analysis**

*The Mushroom Dataset*

Here is a link to the dataset we have decided to use.

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

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?

*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

**Results**

**Conclusions**

**References**

Smithsonian Magazine. (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/

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>