**Mushroom Data Classification**

**CS5950: Machine Learning**

**Final Project**

**By Chris Carlson, Colin MacCreery, and Ben Mechling**

**MUSHROOM DATA**

**The mushroom data was obtained from a UCI data repository (**<https://archive.ics.uci.edu/ml/datasets/Mushroom>**). The data set has 8124 rows and a total of 23 columns. All of the data is categorical. Each value is a single letter representing a value specific to a given attribute (column). The first column “class” indicates whether the mushroom is considered to be poisonous (p) or edible (e). The number of possible values for each attribute varies. The table below shows the possible of each attribute and symbol used to represent it. The largest number of possible values is 12.**

|  |  |
| --- | --- |
| **Attribute** | **Values** |
| class | edible=e, poisonous=p |
| cap.shape | bell=b,conical=c,convex=x,flat=f,knobbed=k,sunken=s |
| cap.surface | fibrous=f,grooves=g,scaly=y,smooth=s |
| cap.color | brown=n,buff=b,cinnamon=c,gray=g,green=r,  pink=p,purple=u,red=e,white=w,yellow=y |
| bruises | bruises=t,no=f |
| odor | almond=a,anise=l,creosote=c,fishy=y,foul=f,  musty=m,none=n,pungent=p,spicy=s |
| gill.attachment | attached=a,descending=d,free=f,notched=n |
| gill.spacing | close=c,crowded=w,distant=d |
| gill.size | broad=b,narrow=n |
| gill.color | 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 | enlarging=e,tapering=t |
| stalk.root | bulbous=b,club=c,cup=u,equal=e,  rhizomorphs=z,rooted=r,missing=? |
| stalk.surface.above.ring | fibrous=f,scaly=y,silky=k,smooth=s |
| stalk.surface.below.ring | fibrous=f,scaly=y,silky=k,smooth=s |
| stalk.color.above.ring | brown=n,buff=b,cinnamon=c,gray=g,orange=o,  pink=p,red=e,white=w,yellow=y |
| stalk.color.below.ring | brown=n,buff=b,cinnamon=c,gray=g,orange=o,  pink=p,red=e,white=w,yellow=y |
| veil.type | partial=p,universal=u |
| veil.color | brown=n,orange=o,white=w,yellow=y |
| ring.number | none=n,one=o,two=t |
| ring.type | cobwebby=c,evanescent=e,flaring=f,large=l,  none=n,pendant=p,sheathing=s,zone=z |
| spore.print.color | black=k,brown=n,buff=b,chocolate=h,green=r,  orange=o,purple=u,white=w,yellow=y |
| population | abundant=a,clustered=c,numerous=n,  scattered=s,several=v,solitary=y |
| habitat | grasses=g,leaves=l,meadows=m,paths=p,  urban=u,waste=w,woods=d |

**Data Observations**

**The attributes values for mushrooms tend to be heavily grouped by the class, making classification somewhat easy.**

**Class**

**The number of elements in each class (edible, poisonous) is roughly equal.**

|  |  |
| --- | --- |
| **Edible** | **Poisonous** |
| 4208 | 3916 |

**Odor**

**The odor class is the most distinctive attribute.**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Class** | **a** | **c** | **f** | **l** | **m** | **n (none)** | **p** | **s** | **y** |
| **Edible** | 400 | 0 | 0 | 400 | 0 | 3408 | 0 | 0 | 0 |
| **Poisonous** | 0 | 192 | 2160 | 0 | 36 | 120 | 256 | 576 | 576 |

**Veil Type**

**The veil type attribute is supposed to have two possible values, but in the data only one value exists. All 8124 rows have value “p” (partial). For Bayesian and regression models this column was excluded from the data.**

**CLASSIFICATION**

**Directly Available Methods**

**Since the data is entirely categorical, numerical methods cannot be directly applied. Methods that can be directly applied are naive Bayesian classification (NBC), classification trees, and random forests.**

**Indirect Methods**

**While implementing a Bayesian classifier, it was observed that the naive Bayesian approach computes a numeric probability estimate value for each element of a test vector. If a dataset was created using these estimates, then numeric methods could be applied as well. Using an estimate of the log-likelihood that the vector is edible given the value of the feature, numeric training and test sets were built. Linear discriminate analysis (LDA), quadratic discriminate analysis (QDA), and logistic regression (Log.Reg.) were applied.**

**Naive Bayesian Classification**

***By Ben Mechling***

***The naïve Bayesian classifier was constructed and***