Mushroom Analysis: A Guide To Predicting Mushroom Edibility

A Demonstration of Classification Implementations

Background:

- Mushrooms possess external features that allow for classification as either poisonous or edible.
- Providing an accurate classification tool can eliminate repetitious, biological testing.
- The University of California, Irvine’s data contains mushrooms’ external features marked as single characters within an array.

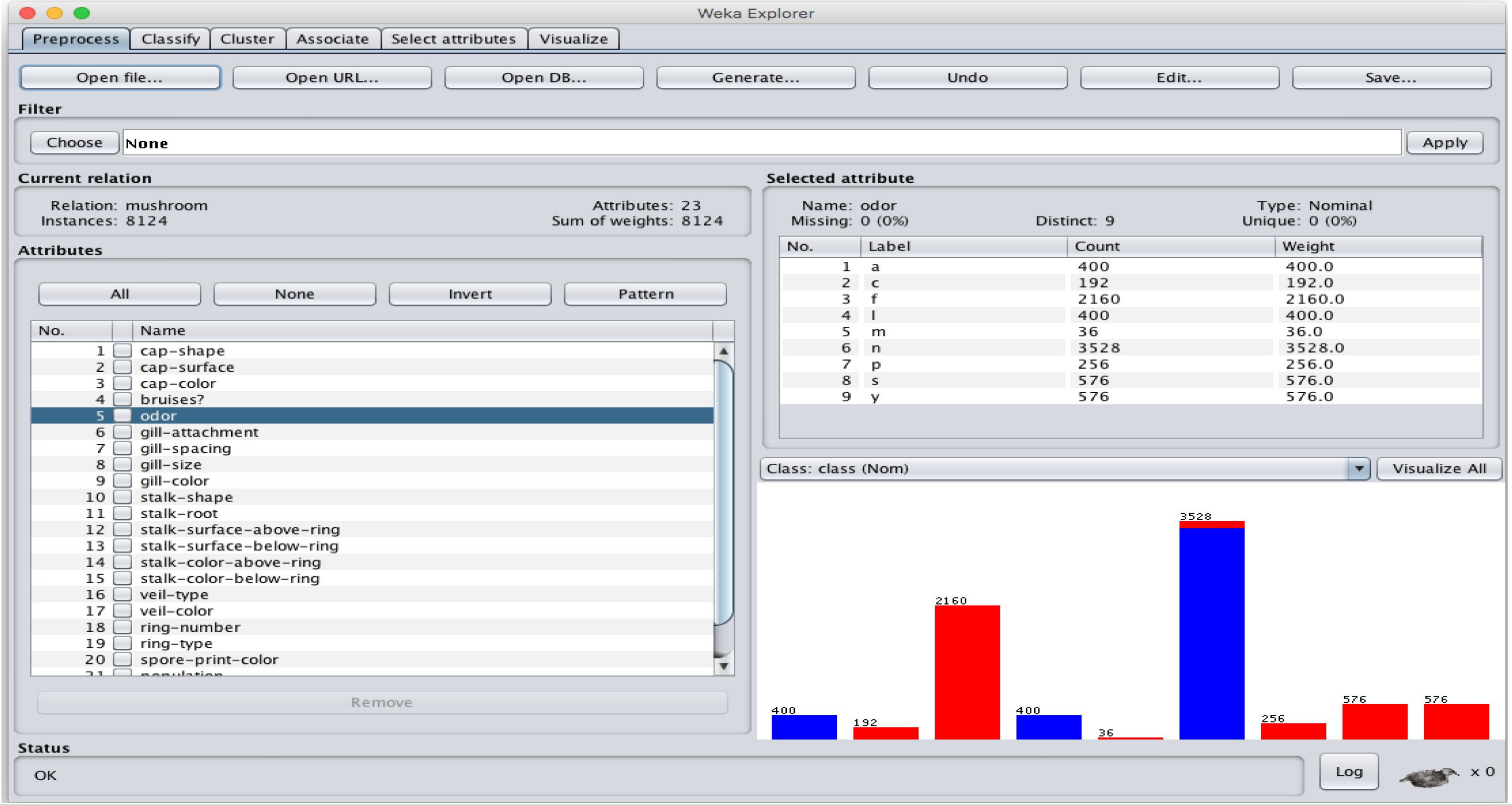
Objective:

- Our group aims to create a classification system that produces *reliable* predictions on the edibility of the mushrooms examined through their trait arrays.
- Our goal is to use the most efficient *and* accurate classification algorithm to achieve our results.
- We also wanted to create a GUI that demonstrated our algorithms and could function in the field.
    - How we handled this goal will be demonstrated further into the presentation.

Weka:

- Java based machine learning toolkit and library
- Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.
- Easy-to-understand GUI and helpful documentation for Java implementation
- From the University of Waikato

Weka: The GUI



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: mushroom
Instances: 8124
Attributes: 23
Sum of weights: 8124
Attributes
All
None
Invert
Pattern
No.
Name
1 cap-shape
2 cap-surface
3 cap-color
4 bruises?
5 odor
6 gill-attachment
7 gill-spacing
8 gill-size
9 gill-color
10 stalk-shape
11 stalk-root
12 stalk-surface-above-ring
13 stalk-surface-below-ring
14 stalk-color-above-ring
15 stalk-color-below-ring
16 veil-type
17 veil-color
18 ring-number
19 ring-type
20 spore-print-color
Remove
Selected attribute
Name: odor
Missing: 0 (0%)
Distinct: 9
Type: Nominal
Unique: 0 (0%)
No. Label Count Weight
1 a 400 400.0
2 c 192 192.0
3 f 2160 2160.0
4 l 400 400.0
5 m 36 36.0
6 n 3528 3528.0
7 p 256 256.0
8 s 576 576.0
9 y 576 576.0
Class: class (Nom)
Visualize All
400
192
2160
400
36
3528
256
576
576
Status
OK
Log
x 0

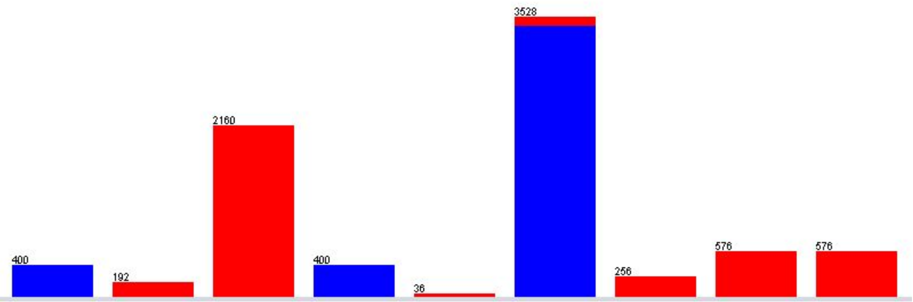
Exploratory Data Analysis (EDA):

- Our data is split into two classes: edible or poisonous
  - This situation requires a binomial classification method.
- The data is a character array of 22 mushroom features with a classification (e or p).
  - For example: x,s,n,f,n,f,w,b,h,t,e,s,s,w,w,p,w,o,e,n,s,g,e
- Which features should we keep and which features should we remove based on uselessness?

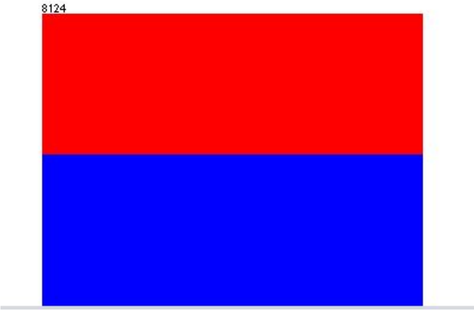
Exploratory Data Analysis (EDA) [continued]:

- We ran 6 different ranker algorithms on our data using the Weka library
- These 6 algorithms returned a unique sequence of features, and each feature was then given a specific rank.

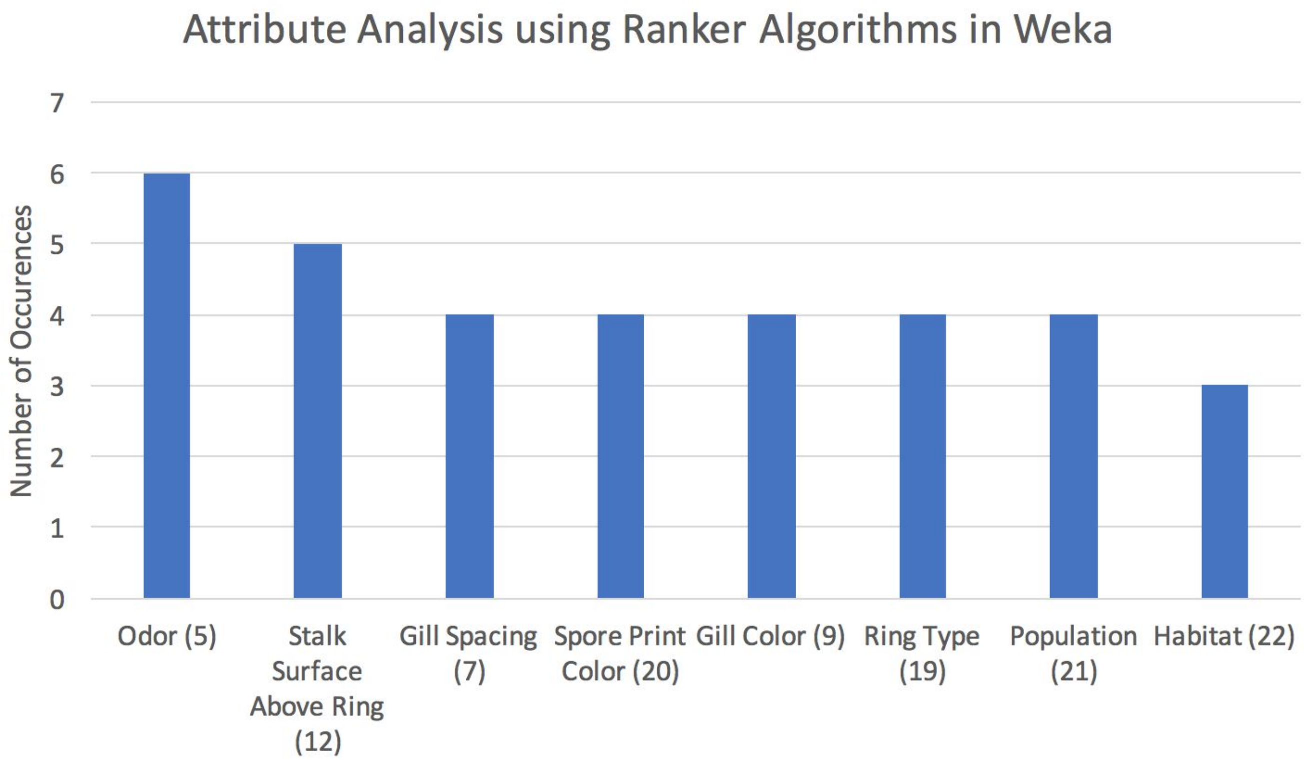
Mushroom Odor Histogram



Veil-Type Histogram



Exploratory Data Analysis (EDA) [continued]:



Implementation #1: Naive Bayes

- We utilized Weka's Naive Bayes classifier because of the reputation and its immediate compatibility with our data's current character array form.
- Naive Bayes proved to be the least complex: simple results with a simple algorithm
- Naive Bayes gives equal weights to features regardless if that feature is significant to the classification regions.

Implementation #2: Logistic Regression

- Logistic Regression is great for predicting binary results.
    - Edible/Poisonous
- Generally limits over-fitting and has low variance from results
- We eliminated 17 features and discovered that logistic regression maintained 100% accuracy, while naive bayes lost accuracy.
    - Those changes in accuracy demonstrate logistic regression's higher potential for dimension reduction.
    - This means we were able to further simplify our data manipulation, while maintaining the same level of reliability (accuracy).
- These benefits are why we chose to pursue this algorithm.

Optimal Algorithm: Logistic Regression

- We chose to pursue logistic regression
  - This is because of base-reliability and with Weka's library utility.
- We abandoned other algorithms such as K-NN, logical ruleset, clustering, and decision trees for reasons such as efficiency and memory.
- Logistic regression returned the highest accuracy in low-dimension conditions.
  - This demonstrates that logistic regression is the best classification algorithm for *this* data.

Difficulties, Solutions and Potential Improvements:

- Problem: Difficulty with using Weka
  - Solution: Fight the learning curve of Weka and convert our data file type to a Weka-supported file type (ARFF).
- Problem: Unsure of Accuracy Reliability
  - Solution: 3-fold and 10-fold cross validation

Visualization:

Logistic Regression Confusion Matrix

Logistic Regression	Edible (Predicted)	Poisonous (Predicted)
Edible (Actual)	1402	0
Poisonous (Actual)	0	1352

- 100% accuracy
- Demonstrates the utility of assigning higher weights to more indicative features.

Naïve Bayes Confusion Matrix

Naïve Bayes	Edible (Predicted)	Poisonous (Predicted)
Edible (Actual)	1402	8
Poisonous (Actual)	21	1331

- 98.95% accuracy
- Demonstrates that equal trait weights can still lead to error simply because statistical significance goes unaccounted for.

Potential Next Steps:

- Collecting our own data
  - Mark each observation with the species so we can return more information
  - Collect data for more than two genera
- Clustering
  - Create a cluster for each genera (Agaricus and Lepiota)
  - Within these, create a cluster for each species (Agaricus: ~300, Lepiota: ~400)

Application: Mushroom Analyzer

- Using our classification algorithms, our group created an android application "Mushroom Analyzer".
- Interested in testing some shrooms! Try it yourself!
- Link: bit.ly/2s9aYtx
- Simply select the features on the mushroom you wish to analyze and our application can determine its edibility.
- Everything is open source! https://github.com/leonardishere/MushroomAnalyzer

Demonstration

- Try our methods for yourself! Here's the download link once more"
    - bit.ly/2s9aYtx
- You can test a mushroom yourself!

Features:
Odor: none
Gill spacing: crowded
Stalk surface: fibrous
Spore print color: black
Population: abundant

Data/Library Resources:

- Weka 3: Data Mining Software in Java
  - http://www.cs.waikato.ac.nz/ml/weka/
- Weka for Android
  - https://github.com/rjmarsan/Weka-for-Android
- University of California, Irvine mushroom data sets
  - https://archive.ics.uci.edu/ml/datasets/Mushroom

Sources

- Mushroom image: http://www.iucnredlist.org/details/75093504/0
- Screenshots from Weka GUI

THANK YOU!