Data Mining Final Project: Mushroom Classification

t

Baskar Dakshinamoorthy & Lauren Foltz

Introduction to Mushrooms

- Growth requirements - energy comes from decaying matter, not sunlight
- Cultivation vs Foraging - Consumption has been going on thousands of years, but cultivation only started in the 1600s.
- Top Mushroom & Truffle Producers:
  - China (5 million tons)
  - Italy (762 thousand tons)
  - United States (391tons) - 41% in Pennsylvania
- Growth
  - First appear as pins, with a cap and stalk

Mushroom Growth

As mushrooms grow, the veil breaks, exposing the gills &releasing spores.

Models:

Text Mining

Naive Bayes

Clustering

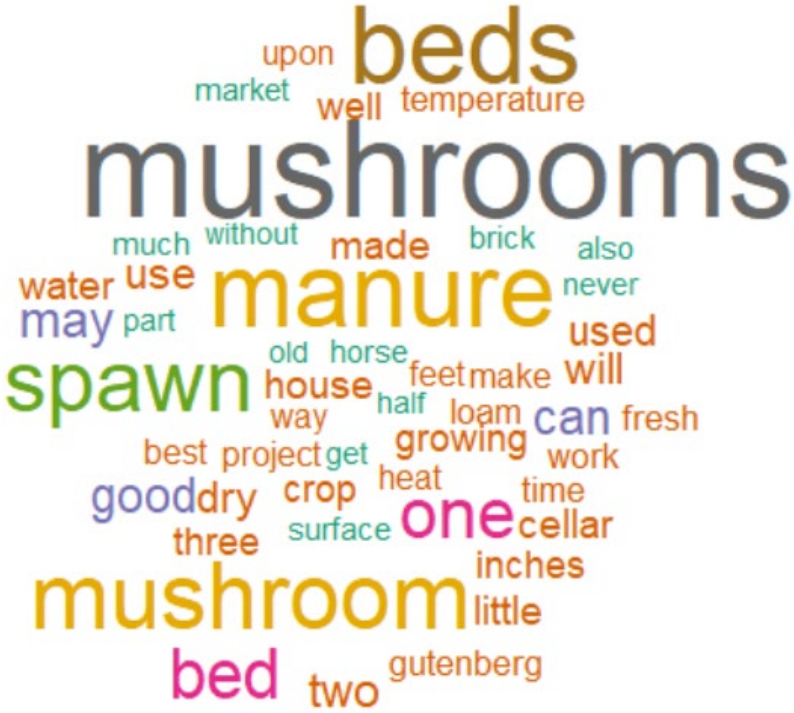
Support Machine Vectors

Text Mining

Mushrooms, How to Grow Them

1892

Beds	451
Manure	420
Spawn	343
Cellar	100
Loam	99
Temperature	92



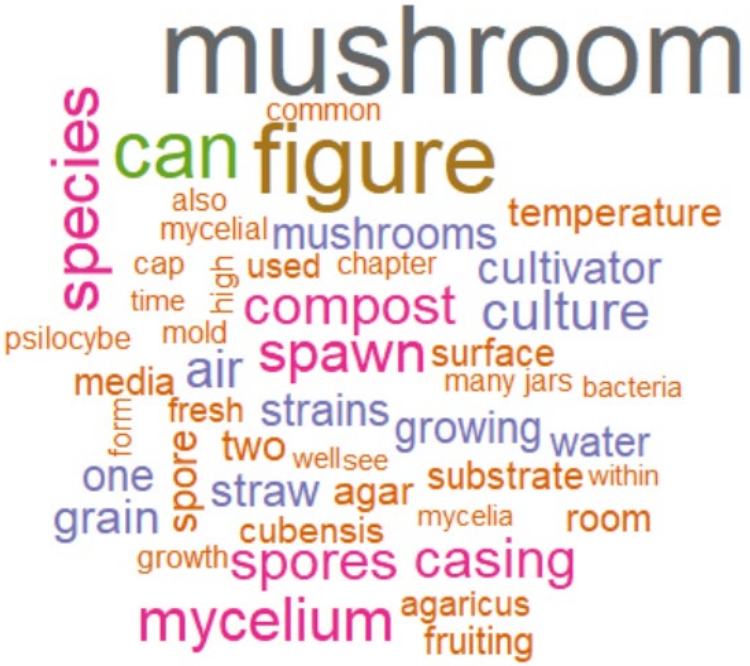
Text Mining

The Mushroom Cultivator

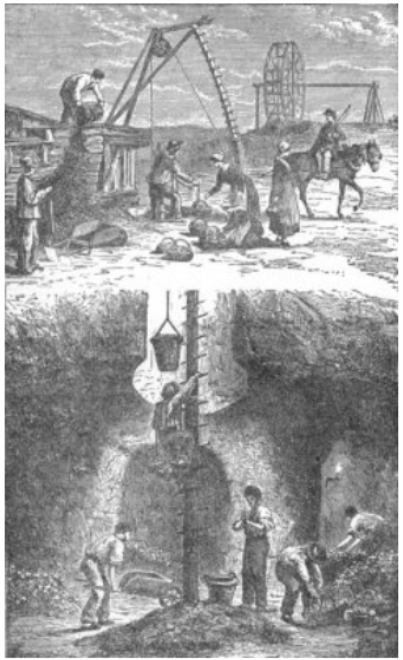
A Practical Guide

1983

	1892	1983
Beds	451	8
Manure	420	8
Spawn	343	82
Cellar	100	0
Loam	99	1
Temperature	92	46



Text Mining



	1892	1983
Beds	451	8
Manure	420	8
Spawn	343	82
Cellar	100	0
Loam	99	1
Temperature	92	46

	1892	1983
Compost	1	76
Straw	37	56
Agar	0	45
Species	7	95
	4708 terms	3199 terms

Naive Bayes

Many poisonous mushrooms were classified as edible.

Model from Naïve Bayes Package:

	TestClassLabels	
nb_prediction	Edible	Poisonous
Edible	1244	113
Poisonous	1	1079

Accuracy: 95.32%
Most information gained from: Odor

Model from e1071 Package:

	TestClassLabels	
NB_e1071_Pred	Edible	Poisonous
Edible	1244	150
Poisonous	1	1042

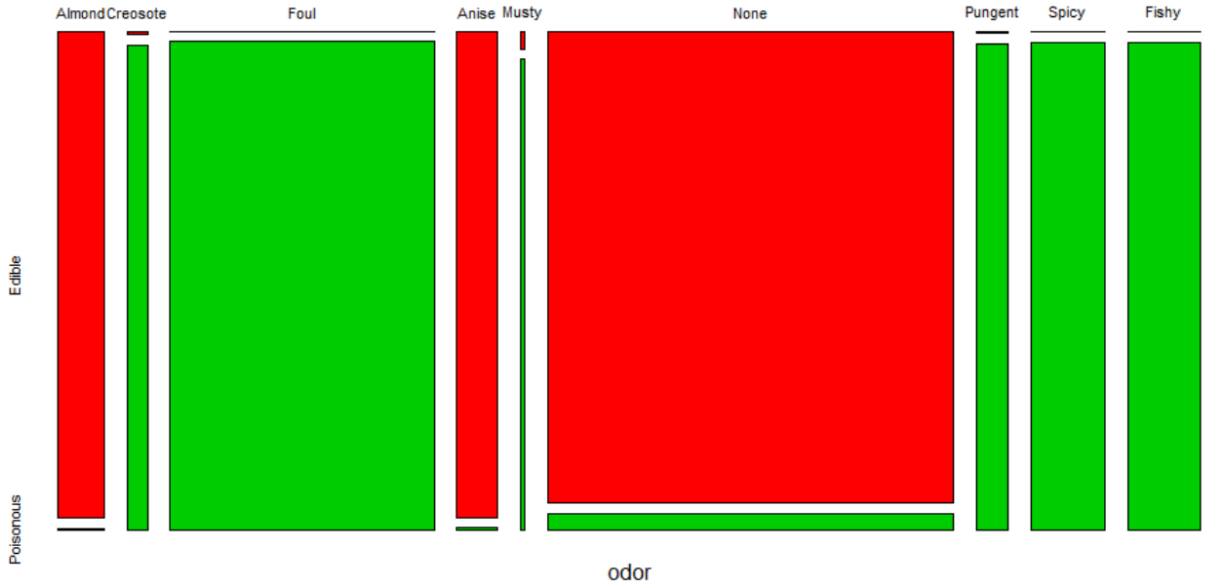
Accuracy: 93.8%
Most information gained from: Odor

Naive Bayes

Odor =

Probably Edible

Strong Predictor

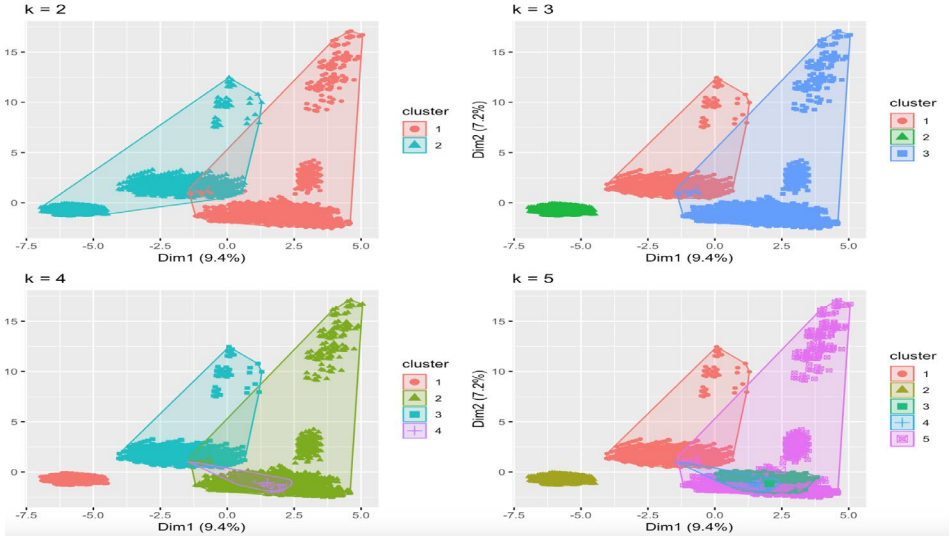


Clustering using k-mode,K-means &Rock Algorithms

Identify optimal clusters for categorical variables

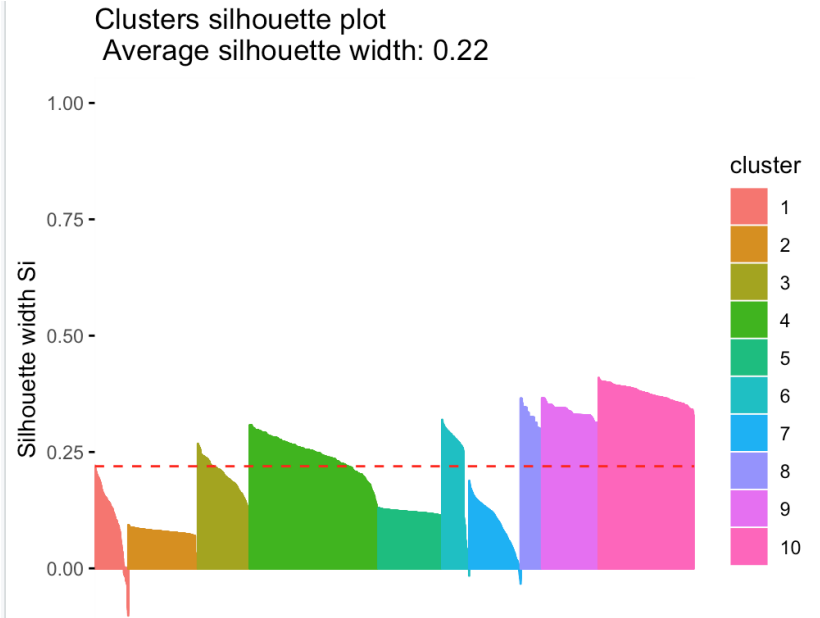
Purity of Clusters for these algorithms

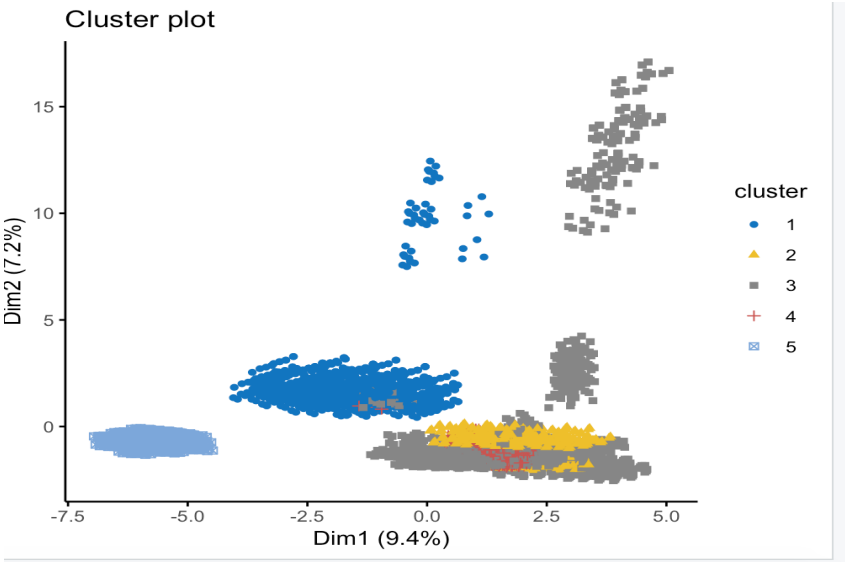
```
> purity.rock
[1] 0.9960611
> purity.kmean
[1] 0.8953717
> purity.kmode
[1] 0.5677006
>
```



DBSCAN and Silhouette Cluster Analysis

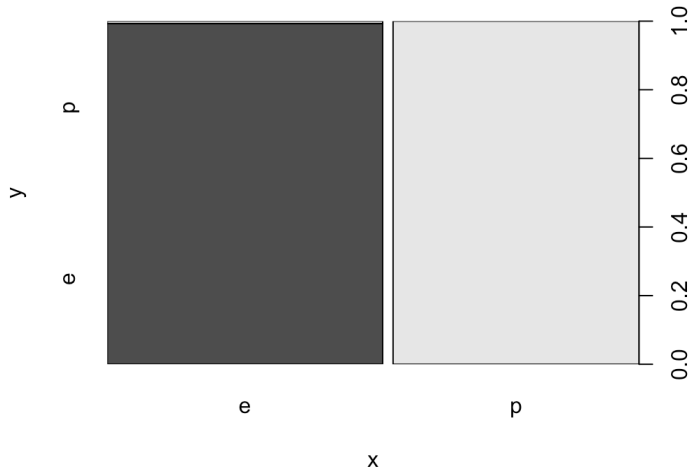
Partition data into groups with similar characteristics using Density based scan and Silhouette Analysis





Support Vector Machines with Radial Basis Kernel

Support Vector Machine Algorithm with Radial Kernel achieved near perfect prediction




```
 Confusion Matrix and Statistics


Prediction          EDIBLE    POISONOUS
        EDIBLE      1278          9
        POISONOUS    0          1150

Accuracy : 0.9963
95% Confidence Interval
```

Summary of Results

Text Mining was used to examine word frequencies within two books from two different time periods.

Naive Bayes agreed with previous Association Rule Mining results that lack of odor was associated with edible mushrooms. It was also in line with previous Decision Tree results that chose Odor as the most important variable.

Clustering using Rock(Robust Clustering using links) achieved a purity of .99 compared to k-means and k-mode.

SVM with radial kernel and C-classification and Random Forest achieved highest prediction rates.

Final Thoughts

DO NOT EAT A MUSHROOM if :

1. Odor is foul
2. Stalk surface above ring is silky
3. Stalk surface below ring is silky
4. Gill size is narrow
5. Spore prints are chocolatey in color