Toxic or Tasty?
— Mushroom Classification

Yueping(Chris) Gu, Autumn Xu, Zhaohua Zheng

Mushroom

Donated on 4/26/1987

From Audobon Society Field Guide; mushrooms described in terms of physical characteristics; classification: poisonous or edible

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Biology	Classification
Feature Type	**# Instances**	**# Features**
Categorical	8124	22

Dataset Information

Additional Information

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). 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 …

SHOW MORE

Has Missing Values?

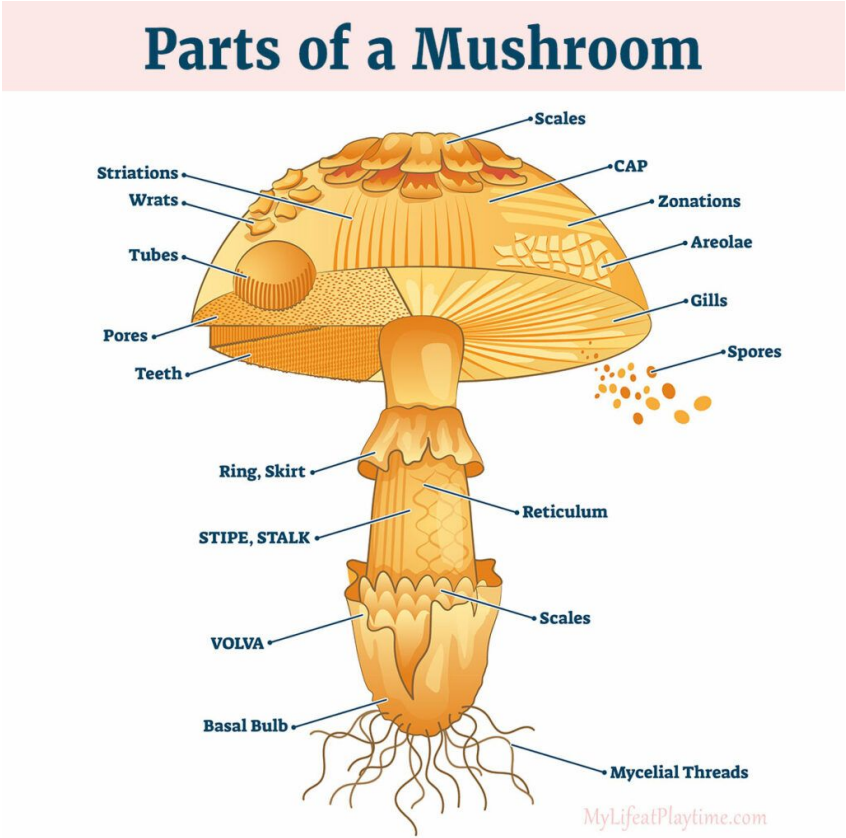
Yes

Variables Table

Variables Table

Variable Name	Role	Type	Description	Units	Missing Values
poisonous	Target	Categorical			no
cap-shape	Feature	Categorical	bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s		no
cap-surface	Feature	Categorical	fibrous=f,grooves=g,scaly=y,smooth=s		no
cap-color	Feature	Binary	brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y		no
bruises	Feature	Categorical	bruises=t,no=f		no
odor	Feature	Categorical	almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s		no
gill-attachment	Feature	Categorical	attached=a,descending=d,free=f,notched=n		no
gill-spacing	Feature	Categorical	close=c,crowded=w,distant=d		no
gill-size	Feature	Categorical	broad=b,narrow=n		no
gill-color	Feature	Categorical	black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y		no

Dataset



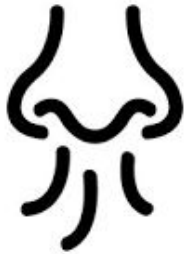
Client

USDA Food Safety and Inspection Service
U.S. DEPARTMENT OF AGRICULTURE

Models

Decision tree
Random Forest
CatBoost
Artificial Neural Network
k-nearest neighbors
Naïve Bayes Classifier

Poisonous/Edible
(true class negative class)



Odor
almond, anise, creosote,
fishy, foul, musty, none,
pungent, spicy

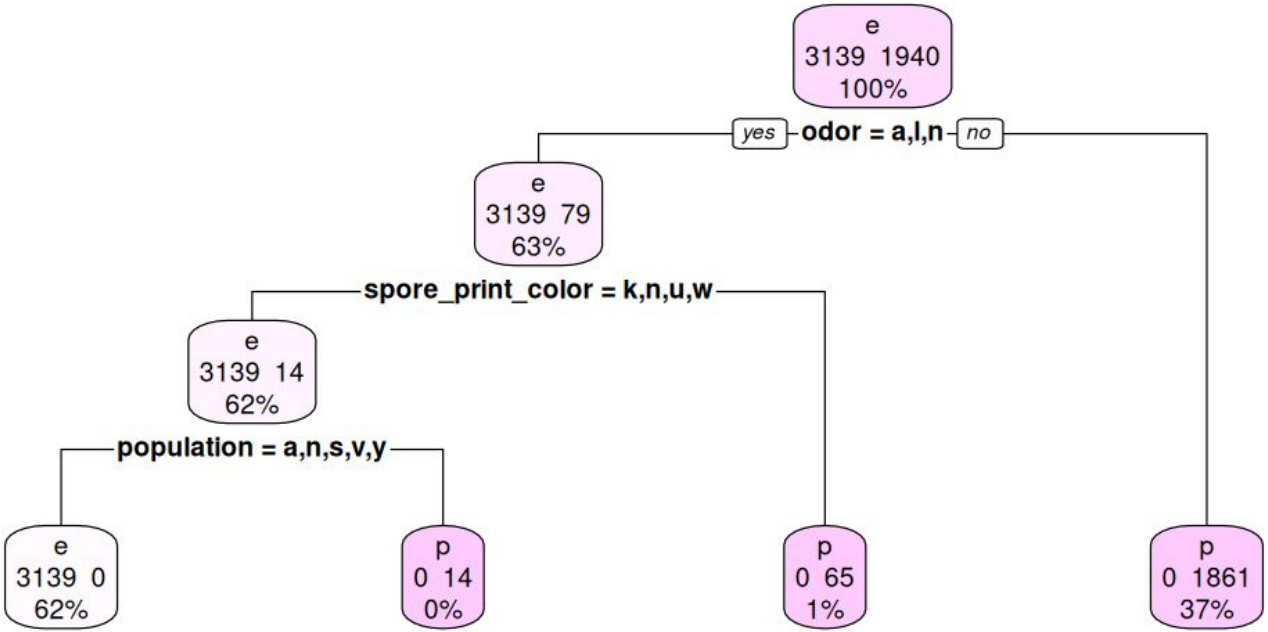
Poisonousness

Poisonous classification: test accuracies

- For all models test accuracy > 0.96

model	test acc (no PCA)	test acc (PCA)
Decision Tree	1.0	0.998
Random Forest	0.9628	1.0
CatBoost	1.0	1.0
ANN	1.0	0.998
KNN	1.0	1.0
Naive Bayes	0.9965	0.9624

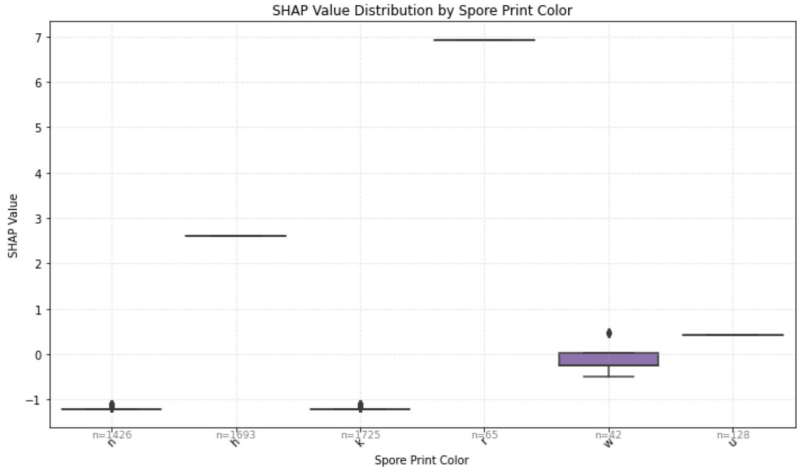
Best Decision Tree Model for Poisonous Classification



Spore print color to predict edibility

result from CatBoost

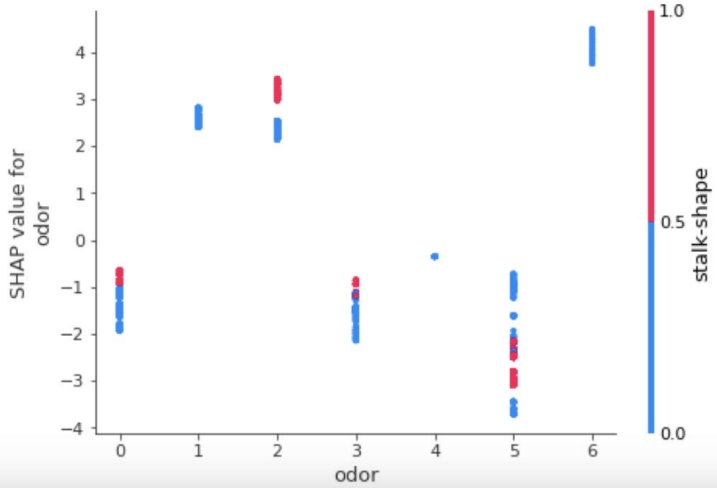
```
# brown → edible
# chocolate → poisonous
# black → edible
# green → strongly poisonous
# white → neutral
# purple → sighly poisonous
```



Odor to predict edibility

result from CatBoost

```
# almond → edible
# creosote → strongly poisonous
# foul → strongly poisonous
# anise → edible
# musty → edible
# none → strongly edible
# pungent → strongly poisonous
# fishy → poisonous
# spicy → poisonous
```



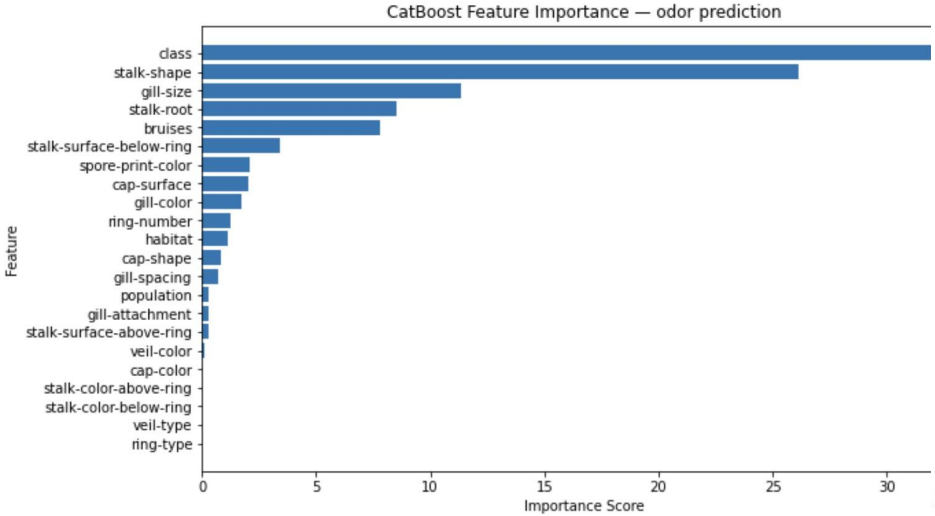
Odor

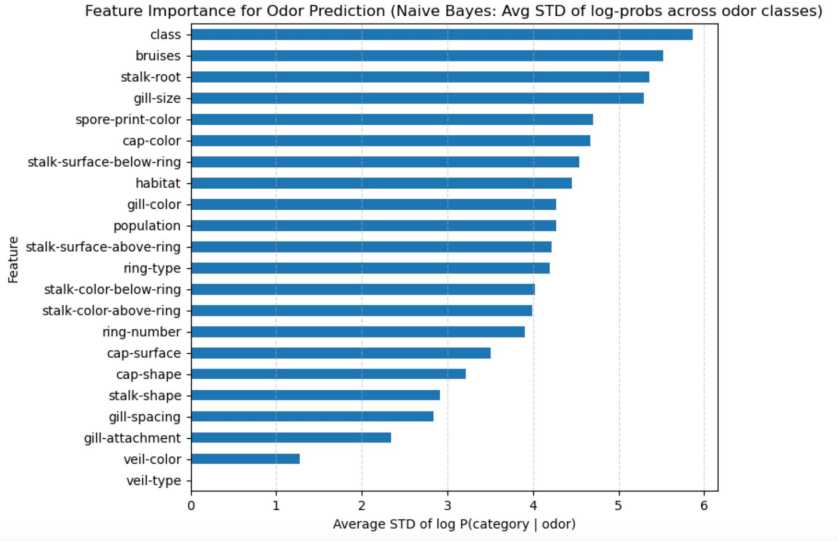
Odor classification: test accuracies

- Test accuracies capped at ~0.922.
- PCA caused slight decreases in test accuracy across most models.

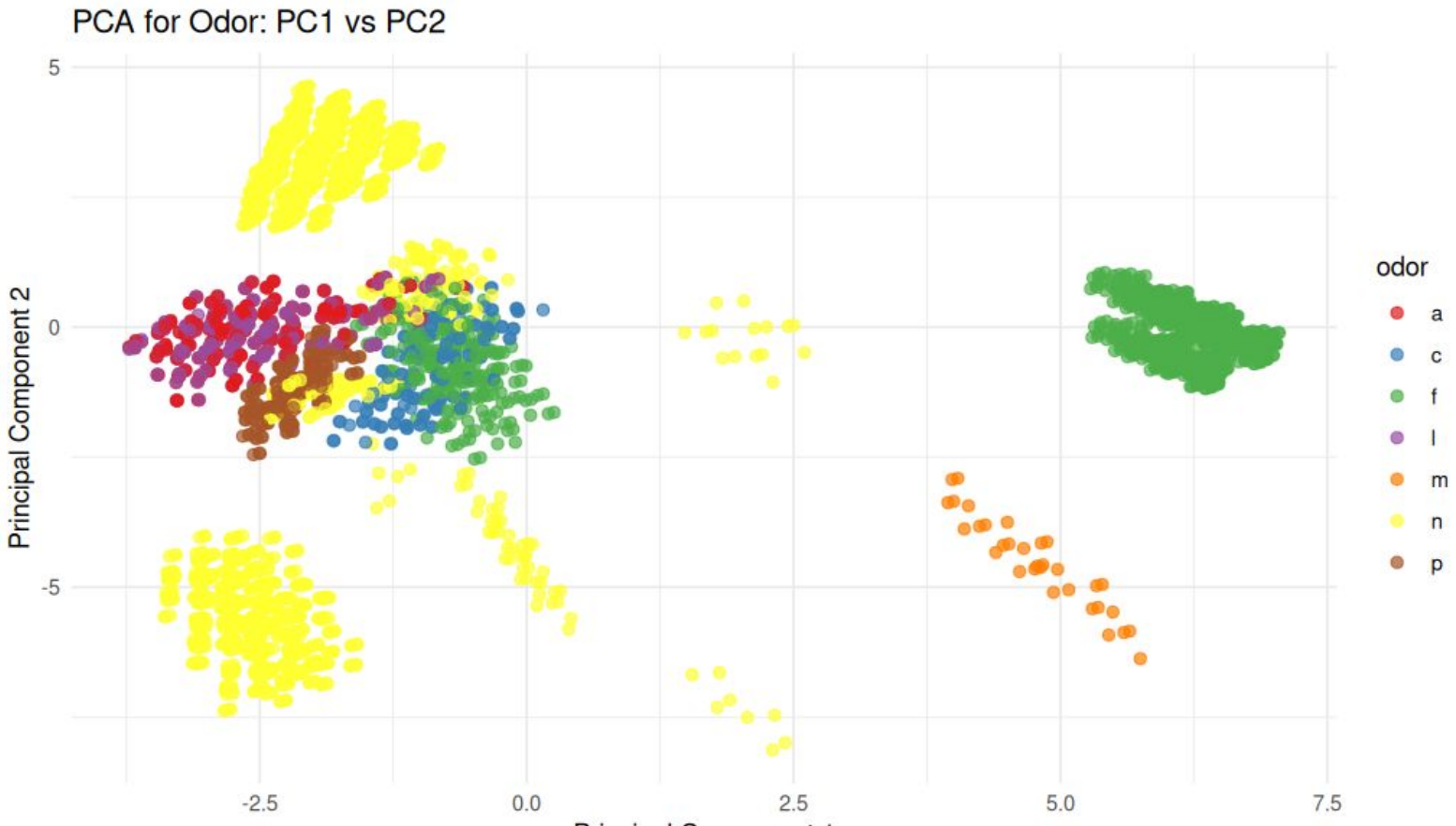
model	test accuracy (no PCA)	test accuracy (PCA)
Decision Tree	0.9221	0.9168
Random Forest	0.8708	0.8673
CatBoost	0.9168	0.8991
ANN	0.9221	0.9221
KNN	0.9133	0.9168
Naive Bayes	0.9080	0.8938

Feature importance for odor prediction





Principal Component Analysis – Dimension Reduction



Recommended actions

More generally,

Consider the effects on runtime and space cost when translating problems between categorical and continuous data

Apply principal component analysis with discretion

Augment with domain-informed feature construction

Benchmark multiple models before optimizing any single one.