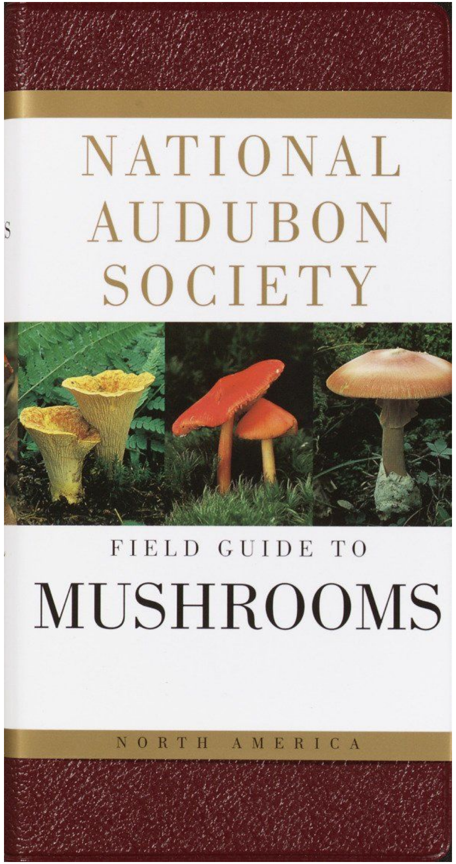
How well does machine learning extract simple logic ???



Eric J. Chan (I.O.D. April 17 2023) – Mini Project 2

Mushroom Data Set

Download: Data Folder, Data Set Description

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

Data Set Characteristics:	Multivariate	**Number of Instances:**	8124	**Area:**	Life
Attribute Characteristics:	Categorical	**Number of Attributes:**	22	**Date Donated**	1987-04-27
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	816892

Source:

Origin:

Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf

Donor:

Jeff Schlimmer (Jeffrey.Schlimmer '@' a.gp.cs.cmu.edu)

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

A hybrid method for extraction of logical rules from data

Włodzisław Duch, Rafał Adamczak, Krzysztof Grąbczewski and Grzegorz Żal
Department of Computer Methods, Nicholas Copernicus University,
Grudziądzka 5, 87-100 Toruń, Poland.
E-mail:duch,raad,kgrabcze,dzezik@phys.uni.torun.pl

Abstract

A hybrid method for extraction of logical rules from data has been developed. The hybrid method is based on a constrained multi-layer perceptron (C-MLP2LN) neural network for selection of relevant features and extraction of preliminary set of logical rules, followed by a search-based optimization method using global minimization technique. Constraints added to the cost function change the MLP network smoothly into a network performing logical operations. The method is applicable for symbolic and continuos features, finding optimal linguistic variables. Results for several medical and other data sets show that such hybrid technique finds very simple and highly accurate rules, frequently giving results that are more accurate than those obtained by any other classifier. Crisp logical rules are found first, followed by fuzzy rules only if the accuracy of the crisp rules is not satisfactory. Comparison with other rule extraction methods shows superiority of the hybrid approach. The method is also applicable in data mining problems.

Keywords

Neural networks, logical rules, optimization, machine learning.

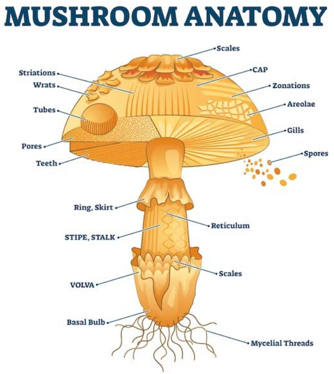
"…Surprisingly, in some applications simple rules proved to be more accurate and were able to generalize better than many machine and neural learning algorithms. Perhaps the main reason for such good performance of logical rules is related to the problem of finding an optimal balance between the flexibility of adaptive models and the danger of overfitting the data…."
Article - July 2000

Many other articles cite use of this dataset as a benchmark for performing logical rule extraction….

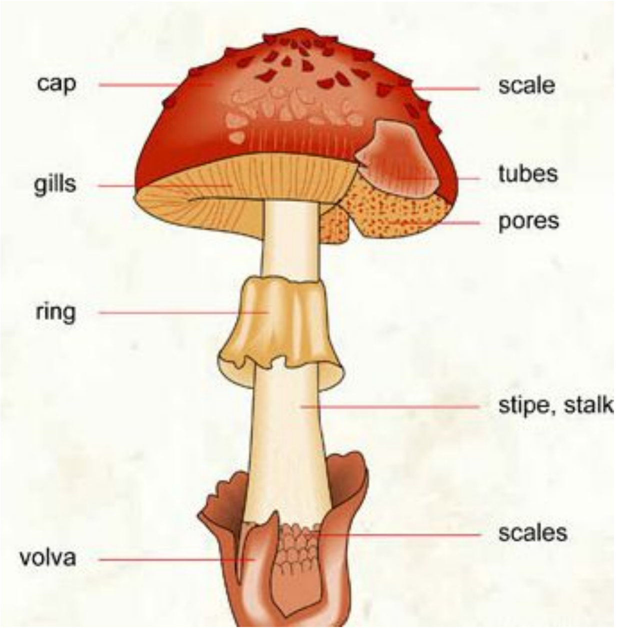
Data Set Information:

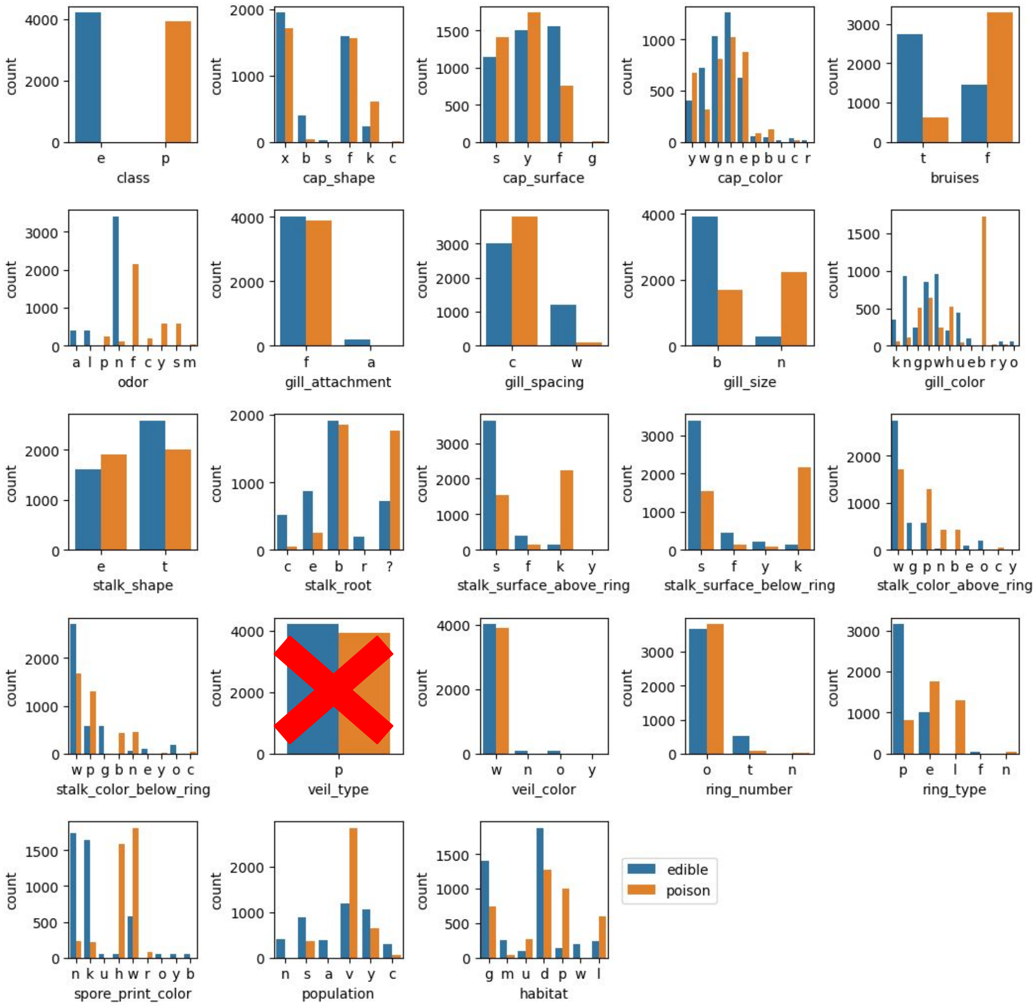
This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous 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.

Attribute Information:



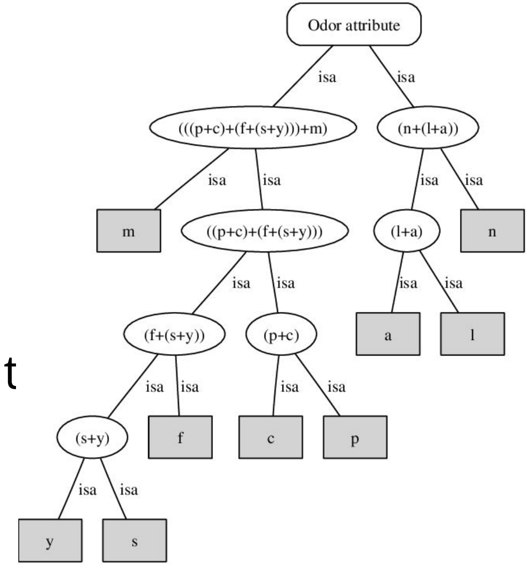
1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?: bruises=t,no=f
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6. gill-attachment: attached=a,descending=d,free=f,notched=n
7. gill-spacing: close=c,crowded=w,distant=d
8. gill-size: broad=b,narrow=n
9. 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
10. stalk-shape: enlarging=e,tapering=t
11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
16. veil-type: partial=p,universal=u
17. veil-color: brown=n,orange=o,white=w,yellow=y
18. ring-number: none=n,one=o,two=t
19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d





1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?: bruises=t,no=f
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6. gill-attachment: attached=a,descending=d,free=f,notched=n
7. gill-spacing: close=c,crowded=w,distant=d
8. gill-size: broad=b,narrow=n
9. 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
10. stalk-shape: enlarging=e,tapering=t
11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
16. veil-type: partial=p,universal=u
17. veil-color: brown=n,orange=o,white=w,yellow=y
18. ring-number: none=n,one=o,two=t
19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Disjunctive logic



- Also known as logical disjunction or inclusive OR.
- A type of logical operation that combines two or more propositions or statements. The resulting compound statement is true if at least one of the individual statement is true.
- i.e. If any of the component statements is true, then the overall statement is true.
- **linear separability** is related to disjunctive logic in the sense that if data points are linearly separable, it means that there exists a **simple rule** (a linear decision boundary) that can accurately classify the data points into different classes.

Image taken from Silvescu, Adrian. (2023). Structural induction: Towards automatic ontology elicitation. Ph.D. Thesis

Disjunctive rules for poisonous mushrooms, from most general to most specific:

Note: The origin of these rules is not clear. At the outset, one may infer human intuition based on experience.

P_1) odor=NOT(almond.OR.anise.OR.none) → 120 poisonous cases missed, 98.52% accuracy

P_2) spore-print-color=green → 48 cases missed, 99.41% accuracy

P_3) odor=none.AND.stalk-surface-below-ring=scaly AND (stalk-color-above-ring=NOT.brown) → 8 cases missed, 99.90% accuracy

P_4) habitat=leaves.AND.cap-color=white → 100% accuracy

Rule **P_4)** may also be **P_4')** population=clustered.AND.cap_color=white → 100% accuracy

These rules involve **6** attributes (out of 22). Rules for edible mushrooms are obtained as negation of the rules given above, for example the rule:

odor=(almond.OR.anise.OR.none).AND.spore-print-color=NOT.green

gives 48 errors, or 99.41% accuracy on the whole dataset.

Several slightly more complex variations on these rules exist, involving other attributes, such as gill_size, gill_spacing, stalk_surface_above_ring, but the rules given above are the simplest we have found.

"Stinkhorn mushroom"

Understanding logical rule P_1

	Predicted Edible (-ve)	Predicted Poison (+ve)
Actually Edible (-ve)	TN: 4208	FP: 0
Actually Poison (+ve)	FN: 120	TP: 3795

number of edible (most frequent) : 4208
proportion edible (bassline accuracy): 0.518
number of poisonous: 3915
Proportion poisonous: 0.482

accuracy of logical rule P_1: 0.9852
recall/sensitivity (tp/(tp+fn)): 0.9693
Precision (tp/(tp+fp)) : 1.0000
f1-score (harmonic mean of precision and recall): 0.9844

- P(m=p):0.482
- P(m=e):0.518
- P(tn)=P(m=e,odor=good)=0.518
- P(fp)=P(m=e,odor=bad)=0.0
- P(tp)=P(m=p,odor=bad)=0.467
- P(fn)=P(m=p,odor=good)=0.015
- P(odor=good|m=p) = 0.015/(0.467+0.015) = 0.0311
- P(m=p|odor=good) = 0.015/(0.518+0.015) = 0.0281
  'fn/(fn+tn) is known as False omission rate'

‘Posterior’ P(m=p|odor=good)=

P(odor=good|m=p) * P(m=p) / P(odor=good)

= 0.0321 * 0.482 / P(m=e,odor=good)+P(m=p,odor=good) =
0.0311 * 0.482 / (0.015+0.518) = 0.0281

P_1) odor=NOT(almond.OR.anise.OR.none) → 120 poisonous cases missed, 98.52% accuracy

P_2) spore-print-color=green → 48 cases missed, 99.41% accuracy

we can see that just by using 'odor' and 'spore_print_color' we can distinguish if a mushroom is poisonous

ie. Simple rule: a foul smelling mushroom with green spores is likely to be poisonous.

odor:
almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s

spore-print-color:
black=k,brown=n,buff=b,chocolate=h,green=r,

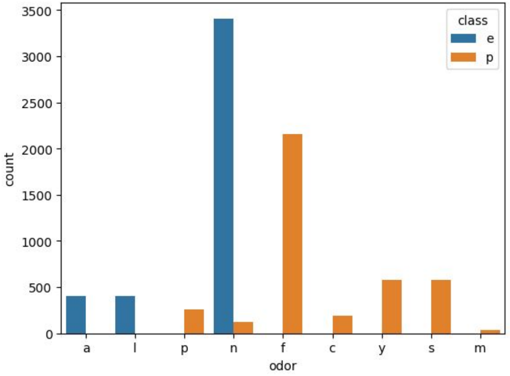
Note: The origin of these rules is not clear. At the outset, one may infer human intuition based on experience.

Investigate the degree to which predictive models may be constructing simple Logical rules ?

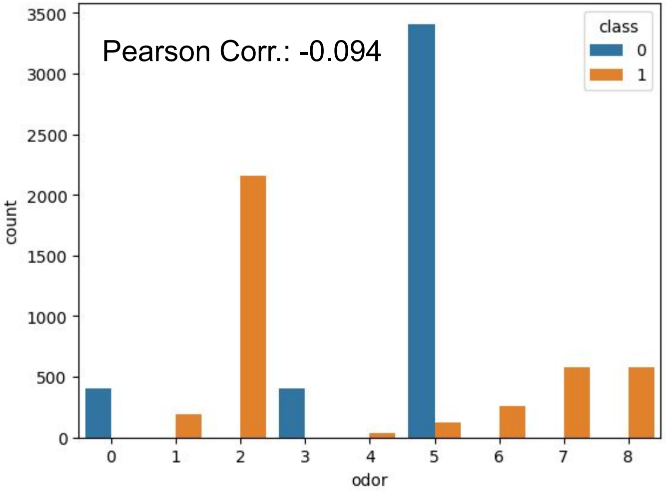
- **Data treatment:** All Nominal encoding and one variant dataset with ordinal encoding only using ‘odor’. 'Veil_type' is removed.
- **Models:** Logistic regression, SVM, Naive Bayes, KNN, Decision Tree
- **Scoring:** Accuracy, F-1
- **Feature Importance:** weights (various regularization penalties), dtc.feature_importances_
- **Feature Selection:** Recursive Feature elimination and SelectFromModel
- **Decision tree attributes:** Depth, Number of nodes, dtc.feature_importances_

Variant dataset with ordinal encoding only using ‘odor’.
Just how important is this ‘odor’ feature??
Can this reasoning be extracted??

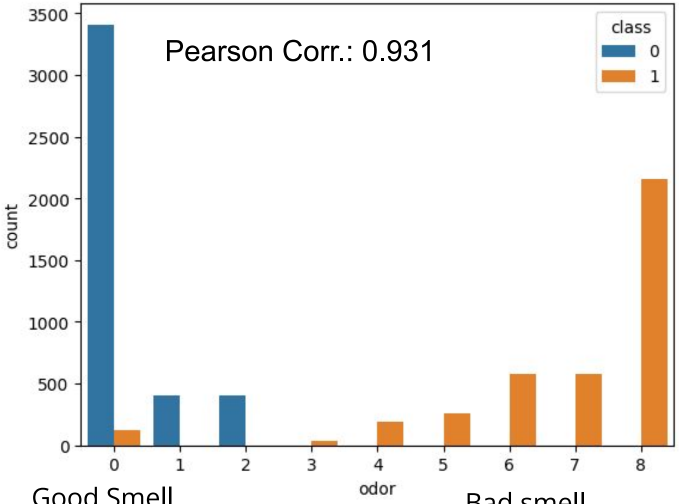
Count plot from initial data



Nominal (alphanumeric)



Ordinal odor order

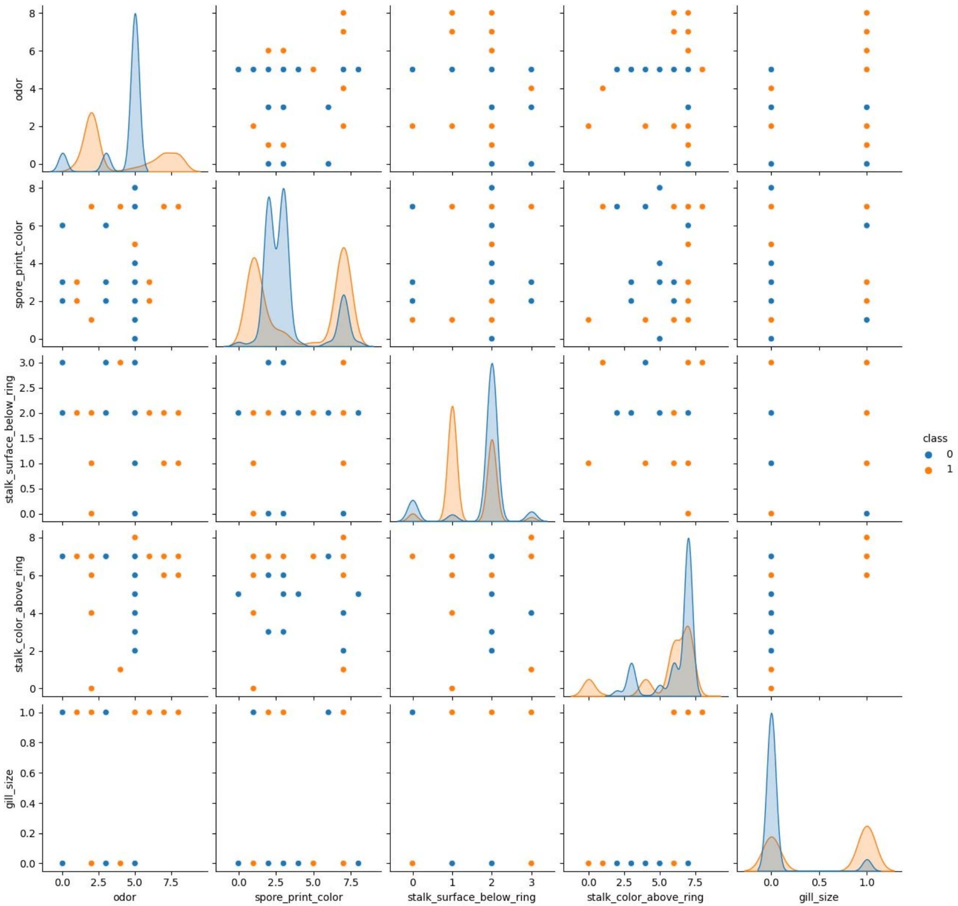


Good Smell
(Least intense)

Bad smell
(more intense)

odor_order=[['none'],['almond'],['anise'],['musty'],['creosote'],['pungent'],['spicy'],['fishy'],['foul']]

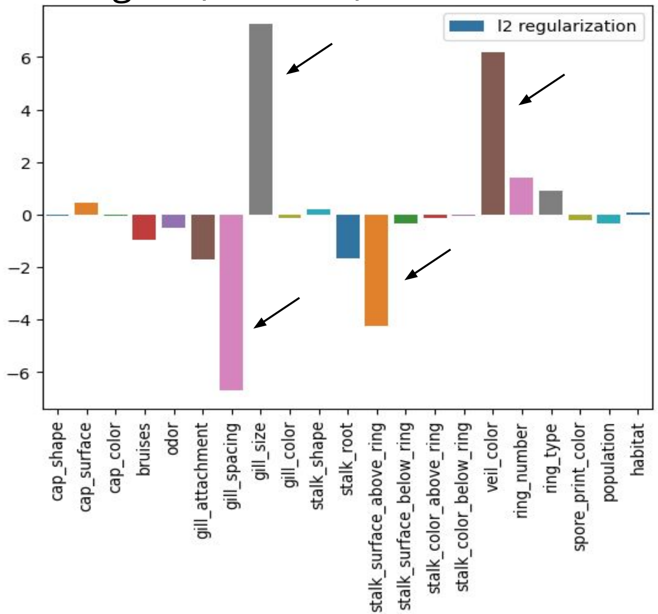
Pair plot of features used for first Three Logical rules and gill size (Blue:edible, Orange:poison)



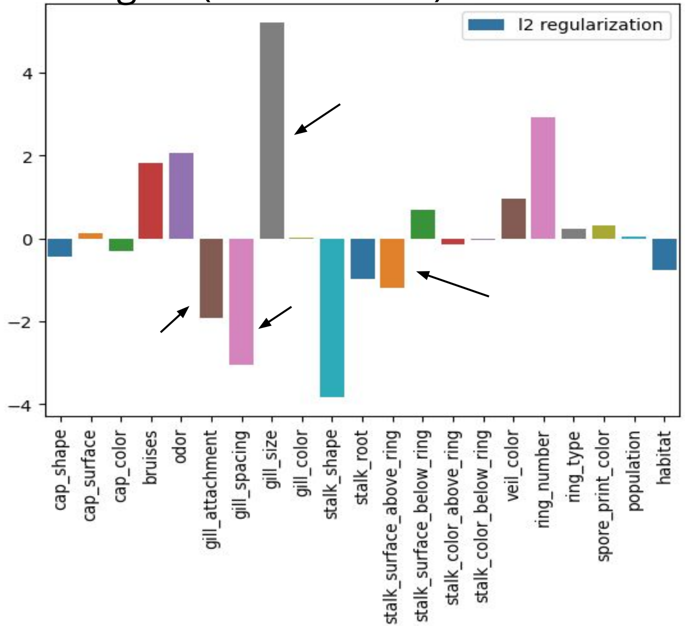
Logistic Regression (L2 regularization)

Type	Encoding	Comment	Solver	Penalty	Test size	Accuracy	Precision	Recall	F1-score	TN	FP	FN	TP
Logistic Regression	nominal	single validation	liblinear	l2	0.2	0.95	0.95	0.94	0.95	758	38	46	783
Logistic Regression	odor ordinal	single validation	liblinear	l2	0.2	0.99	1	0.98	0.99	796	0	15	814

Weights (Nominal)



Weights (odor Ordinal)

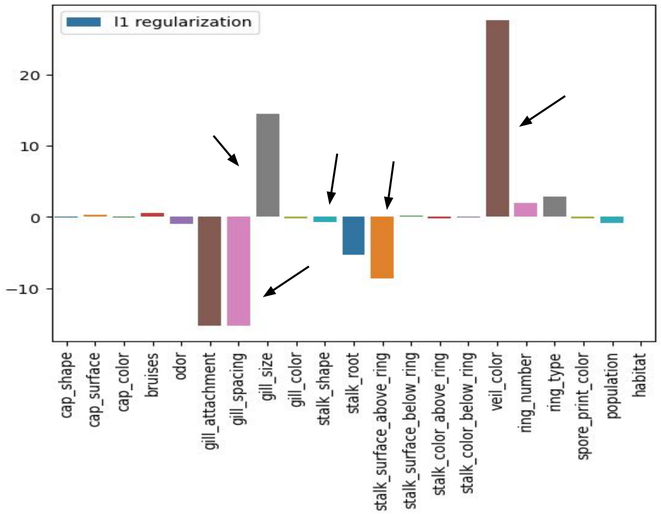


Black arrows indicate best 4 feature from recursive feature elimination

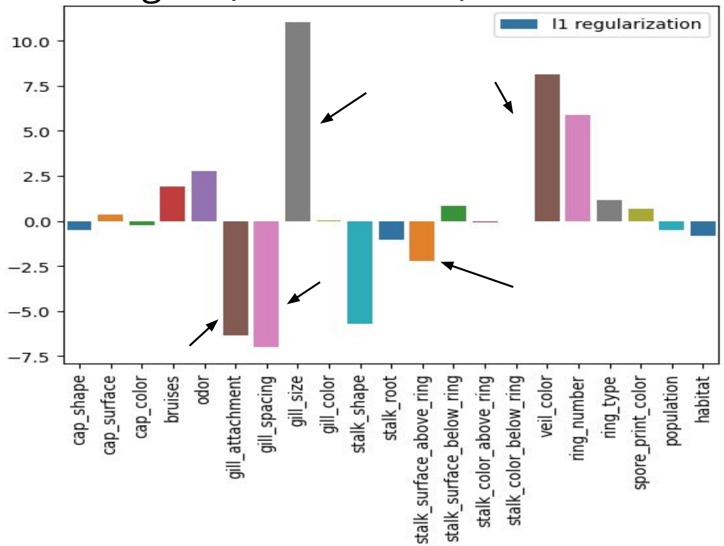
Logistic Regression (L1 regularization)

Type	Encoding	Comment	Solver	Penalty	Test size	Accuracy	Precision	Recall	F1-score	TN	FP	FN	TP
Logistic Regression	nominal	single validation	liblinear	l1	0.2	0.96	0.96	0.96	0.96	764	32	33	796
Logistic Regression	odor ordinal	single validation	liblinear	l1	0.2	0.99	1	0.99	0.99	796	0	11	818

Weights (Nominal)



Weights (odor Ordinal)



Black arrows indicate best 5 feature from recursive feature elimination

Logistic Regression - RFECV

encoding=nominal

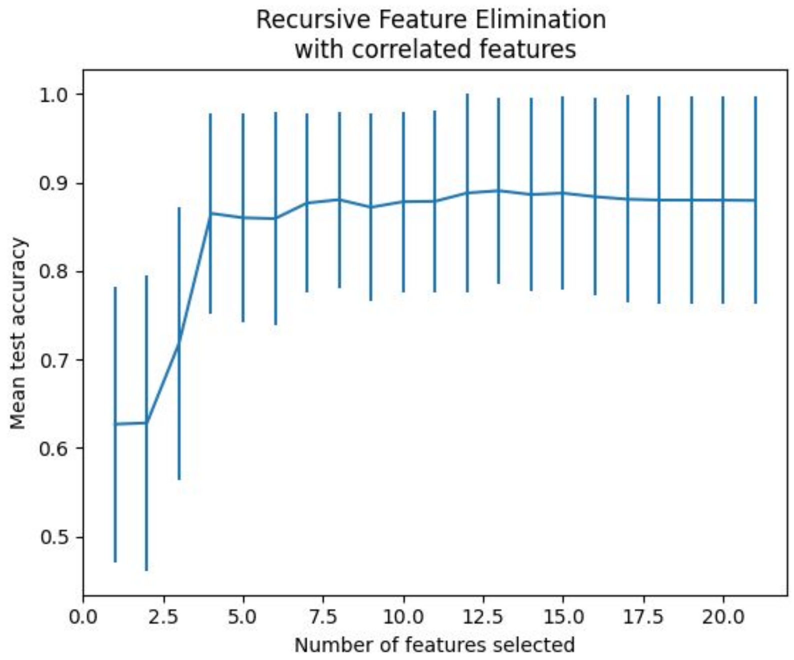
solver='liblinear'

penalty='l1'

scoring="accuracy"

Output: Optimal number of features: 13

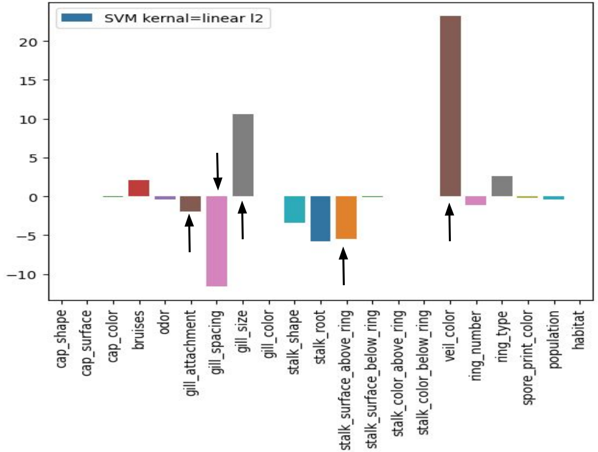
'cap_surface', 'odor', 'gill_attachment', 'gill_spacing', 'gill_size',
'stalk_shape', 'stalk_root', 'stalk_surface_above_ring',
'stalk_surface_below_ring', 'veil_color', 'ring_number', 'ring_type',
'population'



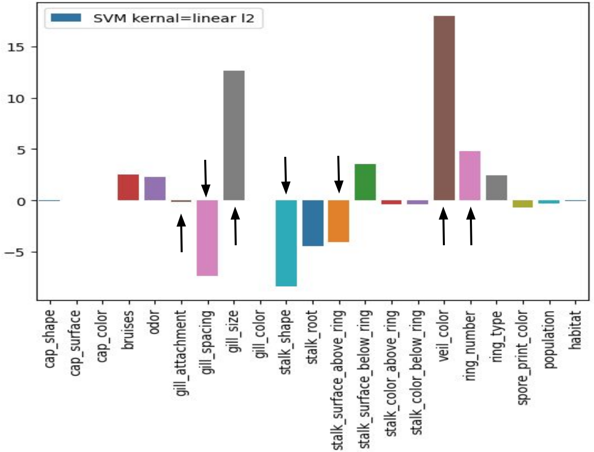
Support Vector Classification (L2)

Type	Encoding	Comment	kernel	reg.	Test size	Accuracy	Precision	Recall	F1-score	TN	FP	FN	TP
SVC	nominal	Best Estimator*	linear	l2 (C=10)	0.2	0.98	0.98	1	0.99	779	17	4	825
SVC	ordinal	Best Estimator*	linear	l2 (C=10)	0.2	1	1	1	1	796	0	0	829
SVC	nominal	single validation	rbf	l2 (C=1)	0.2	1	1	1	1	796	0	0	829
SVC	ordinal	single validation	rbf	l2 (C=1)	0.2	1	1	1	1	796	0	0	829

Weights (Nominal)



Weights (odor Ordinal)




```
* GridSearchCV
svc_params = {
    'C': [1, 10, ],
    'kernel': ['linear','rbf']
}
```

Black arrows indicate best features using LinearSVC and SelectFromModel

Feature Selection using LinearSVC and SelectFromModel

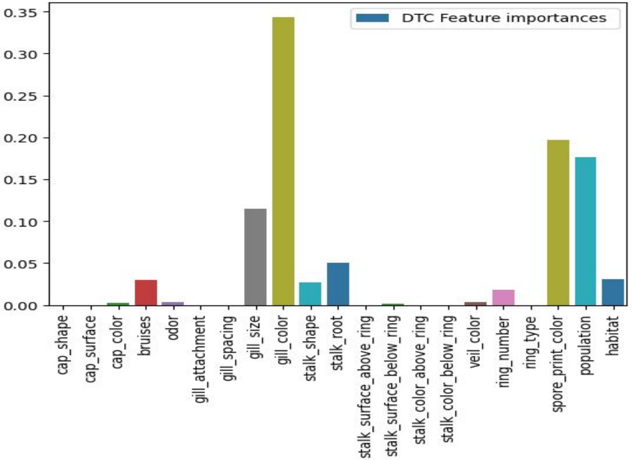
Type	Encoding	Comment	kernel	regularization	Accuracy	number of best features
linearSVC	nominal	dual=False	linear	l1 (C=0.01)	0.94	18
linearSVC	ordinal	dual=False	linear	l1 (C=0.01)	0.99	14
linearSVC	nominal	dual=False	linear	l2 (C=10.0)	0.95	5
linearSVC	ordinal	dual=False	linear	l2 (C=10.0)	1.0	7

Features in bottom two rows correspond with arrows in previous slide

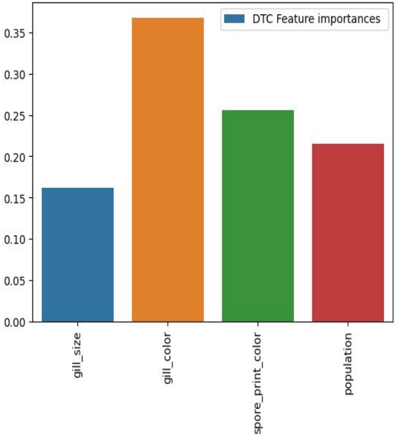
Decision Tree and SelectFromModel(feature importance, scoring='f1')

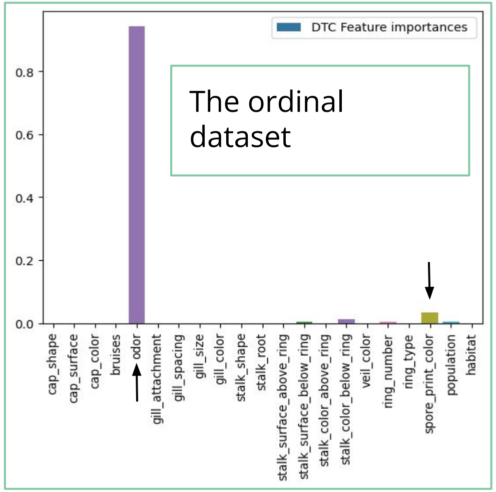
Type	Encoding	Comment	Criterion	Depth,Nodes	Accuracy	Precision	Recall	F1-score	TN	FP	FN	TP
DTC	nominal	GridCV[Gini,entr]	Gini	7,39	1	1	1	1	838	0	0	787
DTC	ordinal*	GridCV[Gini,entr]	Gini	7,17	1	1	1	1	838	0	0	787
SelectFromModel	nominal	X_new = sfm.transform(X)	Gini	11,63	0.99	0.99	0.98	0.99	830	8	15	772
SelectFromModel	nominal	X_new, max_depth=4	Gini	4,25	0.97	0.98	0.96	0.97	805	33	20	767

*The overall crossval score is slightly higher for the ordinal dataset



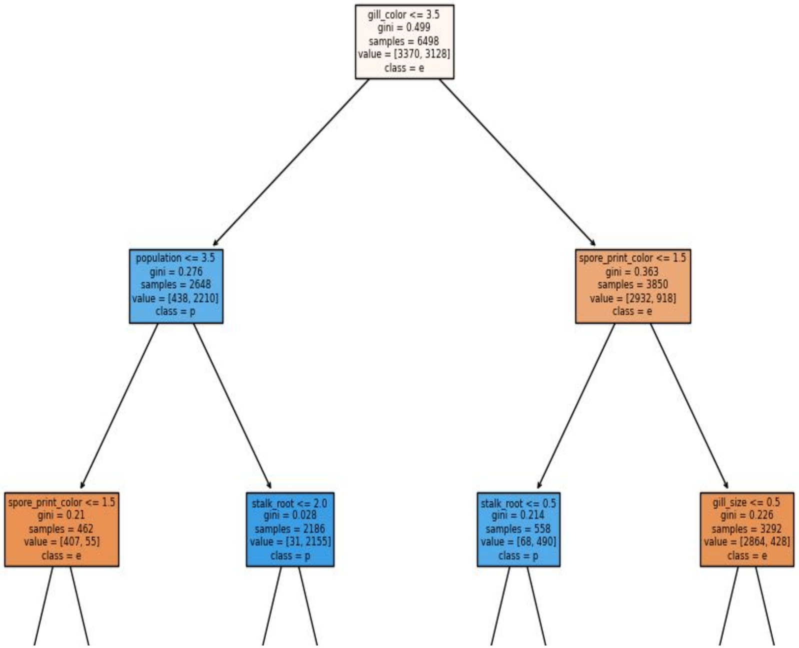
Feature selection using SelectFromModel



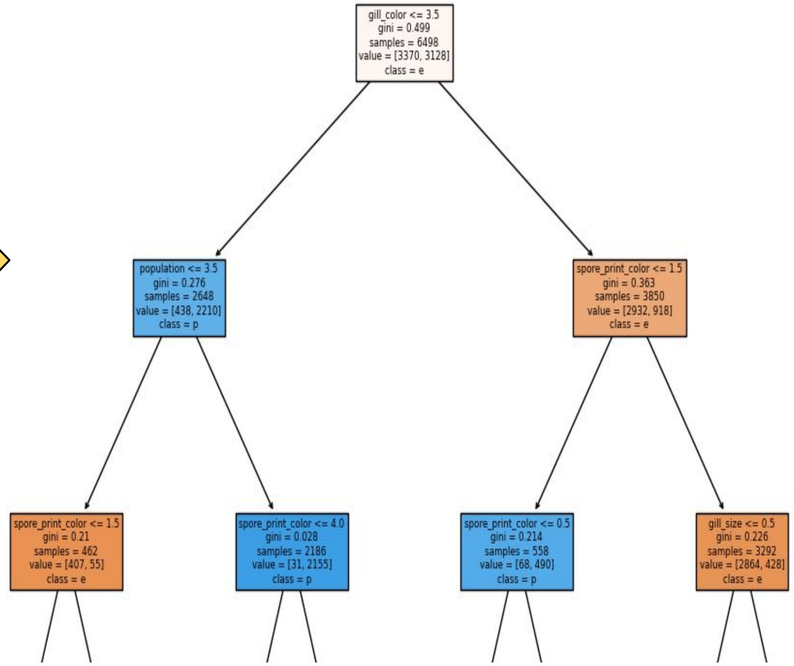


Differences in the initial layers of the trees before and after feature selection.

Features =21



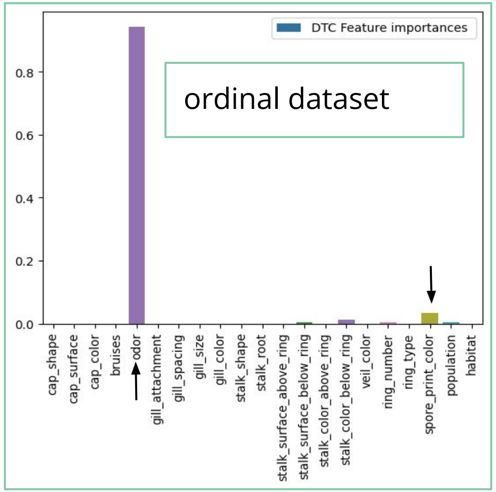
Features =4

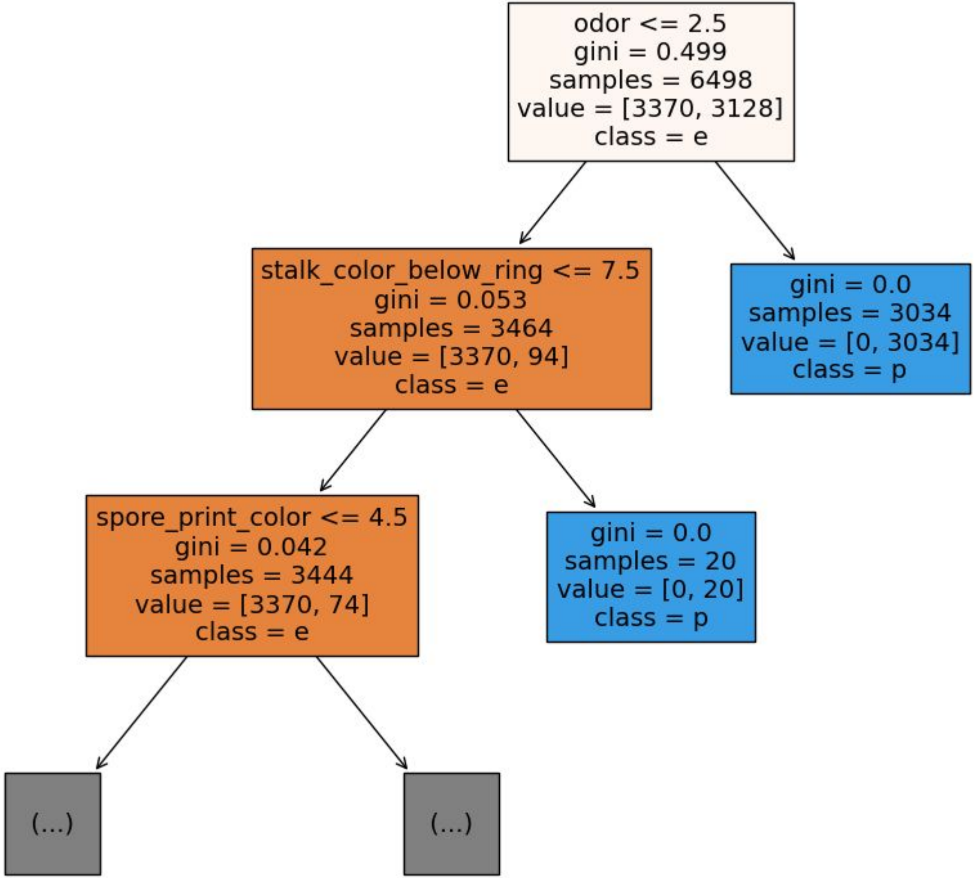


Low GINI : Low impurity ; high GINI : high impurity

Blue : poison ; Orange : edible (intensity of color corresponds with degree of separation)

A Decision Tree as capable of extracting the logical rule, but only from biased data.



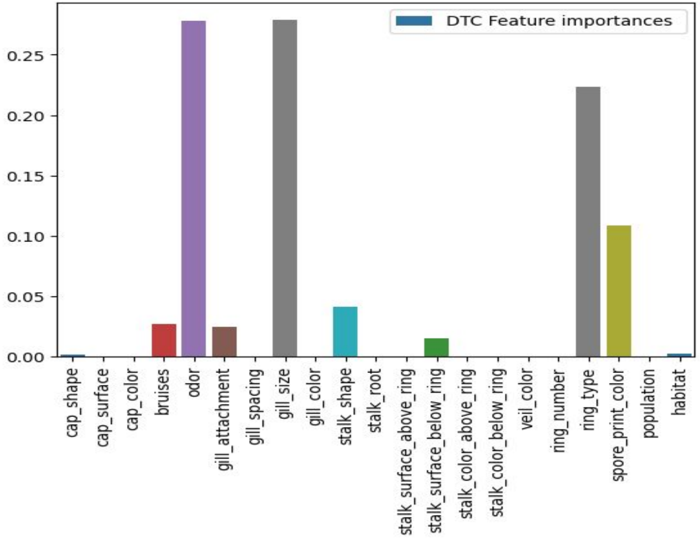


Decision Tree, SelectFromModel GridSearchCV (more params)

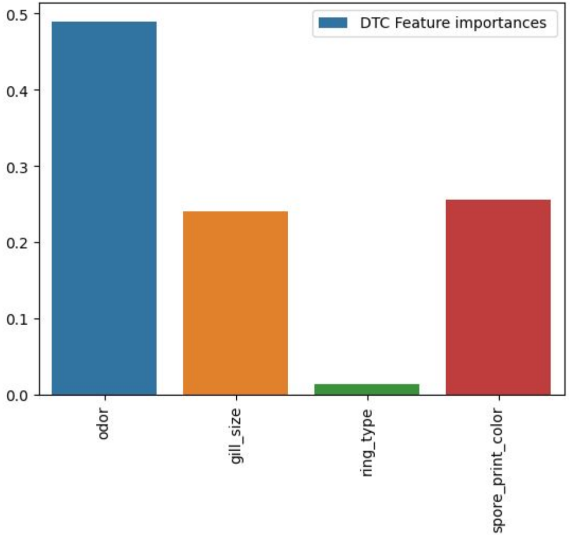
Type	Encoding	Comment	Criterion	Depth,Nodes	Accuracy	Precision	Recall	F1-score	TN	FP	FN	TP
DTC	nominal	GridCV[Many]	entropy	11,47	0.97	-	-	-	-	-	-	-
SelectFromModel	nominal	X_new	Gini	8,39	0.99	0.99	1	0.99	844	9	0	772

```
dtc_params = { 'criterion': ['gini', 'entropy'],
    'max_features': ['auto', 'sqrt', 'log2'],
      'min_samples_split' : [2,5,10],
    'min_samples_leaf': [1, 2, 4],
    'class_weight': [None, "balanced"],
    'min_impurity_decrease': [0.0, 0.1, 0.2]}
```

best_params_: {'class_weight': 'balanced', 'criterion': 'entropy', 'max_features': 'auto', 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 10}

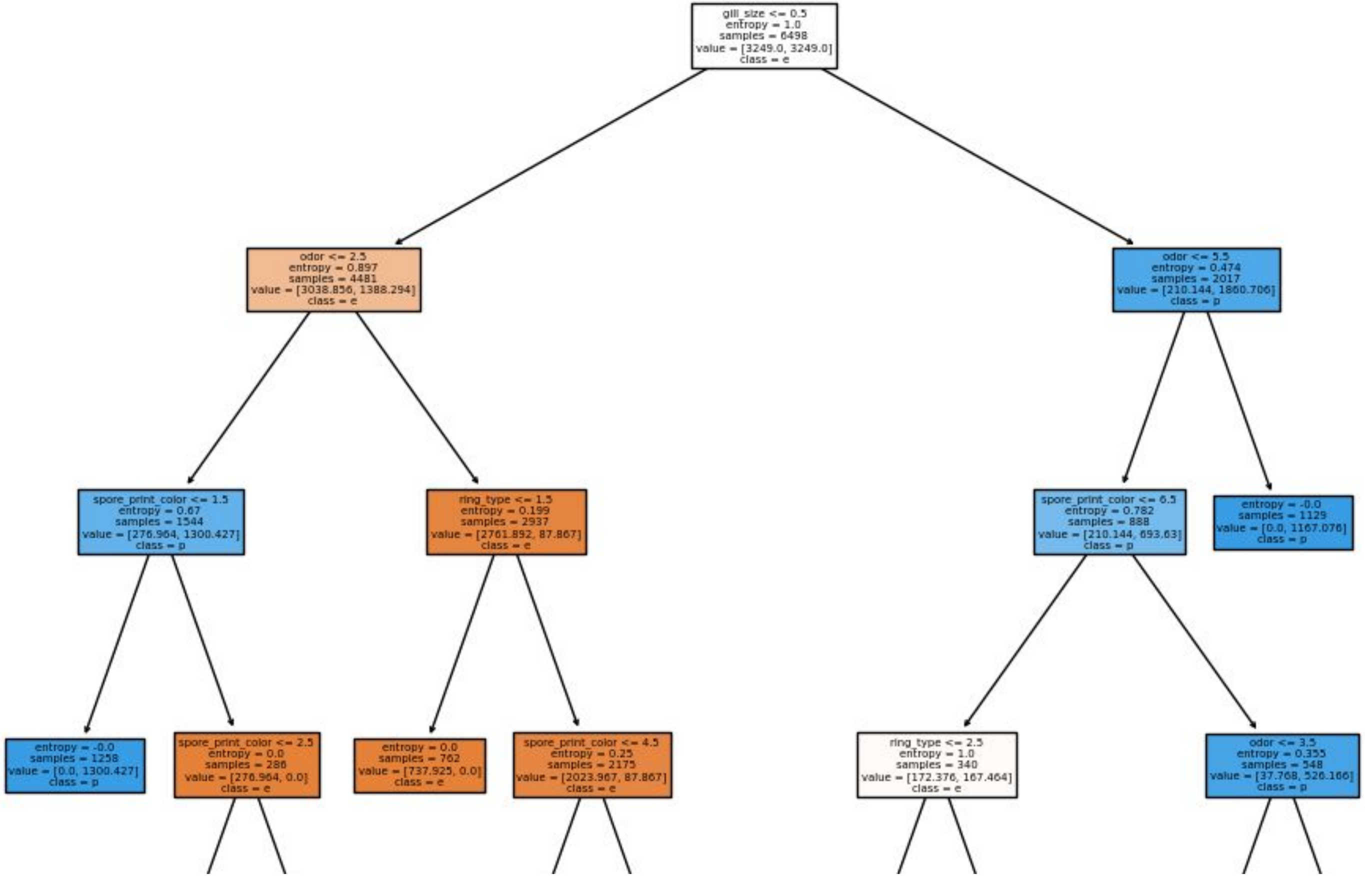


Feature selection using SelectFromModel



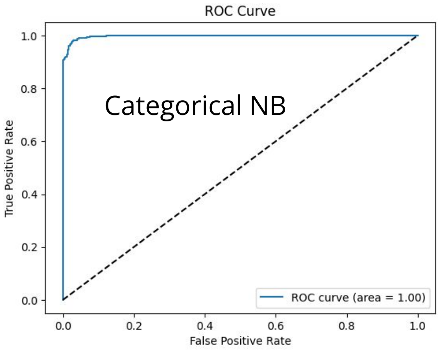
Initial layers of tree after feature selection.
(‘odor’, ‘gill_size’, ’ring_type’, ’spore_print_color’)

High entropy : high impurity
Low entropy : low impurity



Naive Bayes and KNN

Type	Encoding	Comment	Penalty	Test size	Accuracy	Precision	Recall	F1-score	TN	FP	FN	TP
GaussianNB	nominal	single validation	-	0.2	0.91	0.91	0.92	0.91	716	80	64	765
GaussianNB	ordinal	single validation	-	0.2	0.98	0.98	0.97	0.98	783	13	26	803
CategoricalNB	nominal	single validation	alpha=1	0.2	0.96	0.99	0.92	0.95	847	6	64	708
CategoricalNB	ordinal	single validation	alpha=1	0.2	0.96	0.99	0.92	0.95	847	6	64	708
KNN	nominal	*best K = 1	-	0.2	1	1	1	1	796	0	0	829
KNN	ordinal	*best K = 1,2,3...	-	0.2	1	1	1	1	796	0	0	829



*If nominal dataset was used then acc=1.0 only for K=1, acc=0.99 for K>1. for ordinal dataset acc=1.0 for many K>1

Conclusions

- Unclear if the logical rule for odor in the mushroom dataset comes from human intuition. However, it can be inferred that the logical rule is based on empirical observations of the physical characteristics of mushrooms and their edibility.
- Odor feature is one of the most important features for predicting whether a mushroom is poisonous or edible, according to other feature importance analysis results (based on Random forest and gradient boosting see https://towardsdatascience.com/feature-importance-to-predict-mushrooms-edibility-in-python-5aa133fea3f8).
- Using the less-sophisticated ML approaches the odor feature does not rank high in feature importance. Gill size was often ranked higher.
- A decision tree is able to identify odor as a feature of importance. This was striking for when odor was encoded as ordinal. To identify a similar relation from the nominal dataset, a more sophisticated screening with decision tree models using a multi-parameter grid search and cross-validation was necessary.

Spare slides below

As a language model AI, I cannot have opinions or beliefs, but I can give you some suggestions on how to classify the Mushroom dataset available on the UCI website. The Mushroom dataset is a well-known dataset in the machine learning community used for classification. It contains 8124 samples of mushrooms with 23 features, including the cap shape, cap surface, cap color, bruises, odor, gill attachment, gill spacing, gill size, gill color, stalk shape, stalk root, stalk surface above ring, stalk surface below ring, stalk color above ring, stalk color below ring, veil type, veil color, ring number, ring type, spore print color, population, habitat, and the target variable indicating whether the mushroom is edible or poisonous.

Here are some suggestions on how to classify the Mushroom dataset:

1. Perform data exploration and visualization to understand the relationship between the features and the target variable.

2. Preprocess the data by encoding the categorical features using one-hot encoding or ordinal encoding.

3. Split the data into training and testing sets.

4. Train several classification models such as Decision Trees, Random Forest, Naive Bayes, Logistic Regression, Support Vector Machines, and Neural Networks on the training data.

5. Evaluate the performance of the models using metrics such as accuracy, precision, recall, and F1-score on the testing data.

6. Select the best-performing model based on the evaluation metrics and use it to make predictions on new data.

Additionally, you can also perform feature selection and engineering to identify the most important features to improve the performance of the models. Ensuring that the training dataset is well balanced with edible and poisonous mushrooms can also help build better models. It is important to note that the Mushroom dataset is a binary classification problem, and you must be very careful when interpreting the results.

Disjunctive Logic has to do with linear seprability and the ability to identity simple rule from complex highly diversified emmergnt phenomena


Description
For this mini-project you will apply predictive modelling on an area of your choice. Choose a dataset or select a question and identify and gather the data you need. Follow the steps presented in the course so far:

a) Perform EDA on the data
b) Select the outcome/response variable
c) Select features (optionally perform feature engineering)
d) Select approaches/models
e) Apply the models and evaluate them

Present from slides in 8-12 minutes on Monday 17 April.

Rubric:
Explanation of the business context behind the dataset
Quality of presentation - engaging, well structured, not too short or too long
Quality of the notebook - code well documented, runs correctly