Report

Project: 0261 Project 3

2024

Contents

Important Note

This project is for educational purposes only.

Picking and eating mushrooms without knowledge can be really dangerous and it is not endorsed with this project.

People should always ask a professional mycologist or expert in the field before picking, collecting and eating mushrooms.

The authors are not and will not be responsible for any damage done by using this information.

0.1 Dataset Information

0.1.1 Dataset Source

The information and dataset were gathered from the following sources.

- UCI ML Repository

Dataset Features

The dataset includes descriptions of samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family.*Mushroom* (1987) Each species is identified as definitely edible, definitely poisonous, or of unknown edibility, which was combined with the poisonous class.

The dataset contains the following features and descriptors.

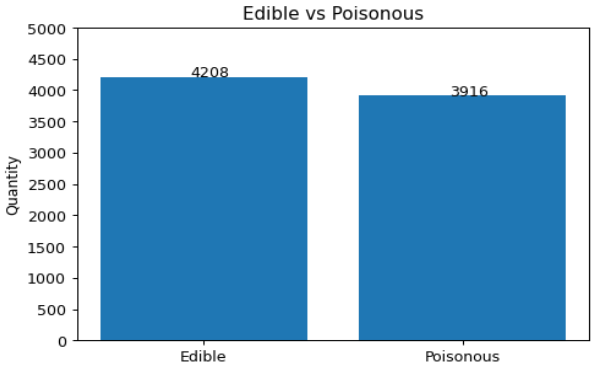
No.	Feature	Description
1	cap_shape	bell=b,conical=c,convex=x,flat=f,knobbed=k,sunken=s
2	cap_surface	fibrous=f,grooves=g,scaly=y,smooth=s
3	cap_color	brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p, purple=u,red=e,white=w,yellow=y
4	bruise	bruises=t,no=f
5	odor	almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m, 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

0.2 EDA

0.2.1 Class Label Analysis

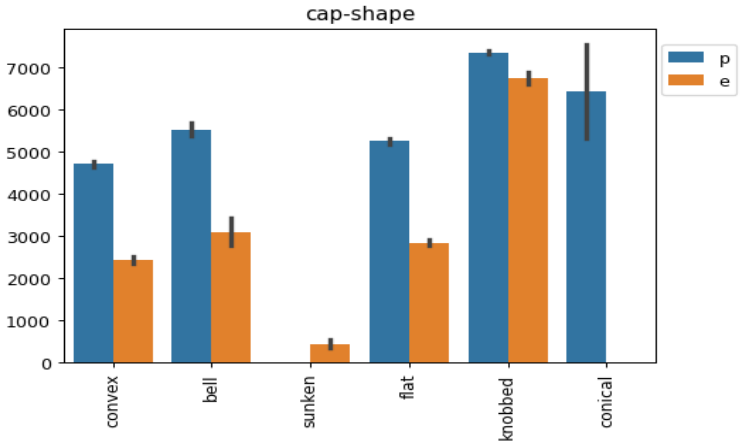
The class labels are balanced. Edible has around 51.8 % and Poisonous has around 48.2 % from a total of 8,123 rows of information.

Class	Percentage
edible	51.80 %
poisonous	48.20 %



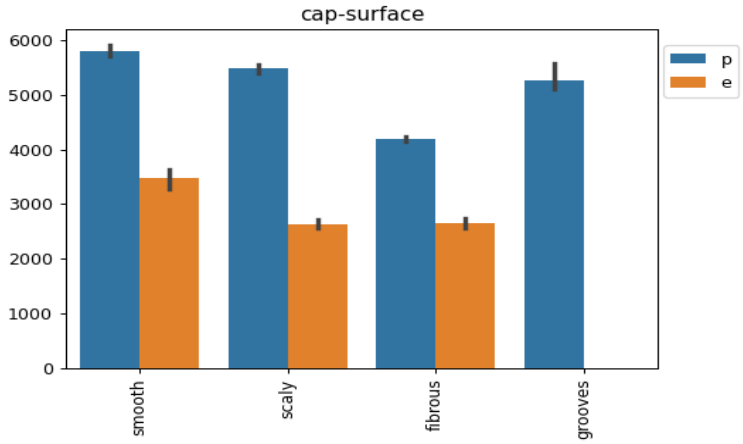
0.2.2 Feature Analysis

The Poisonous class has a distinguished conical cap-shape that edibles don't seem to have.
There is a small percentage of edible mushrooms that have a shrunken cap shape. The amount is so small that it is not a good indicator for this class.
Other indicators will need to be taken into account.

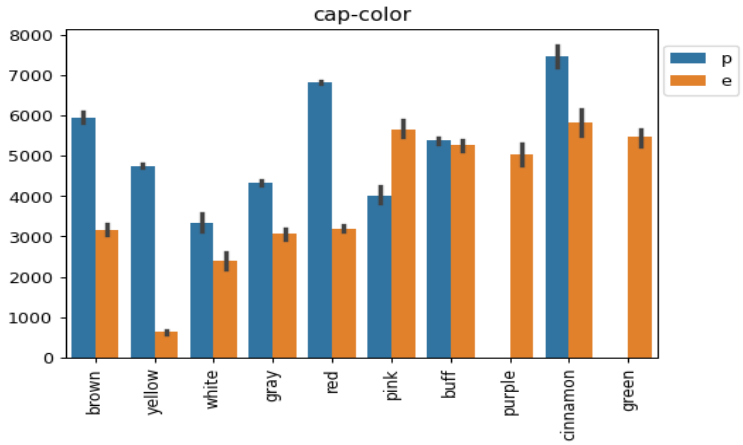


There are indicators on the cap surface that are more prevalent on the poisonous class, but still appear on the edible class.
Having grooves on the cap surface seems to be a strong indicator that the mushroom is poisonous. No edible mushroom seems to have grooves on this dataset.

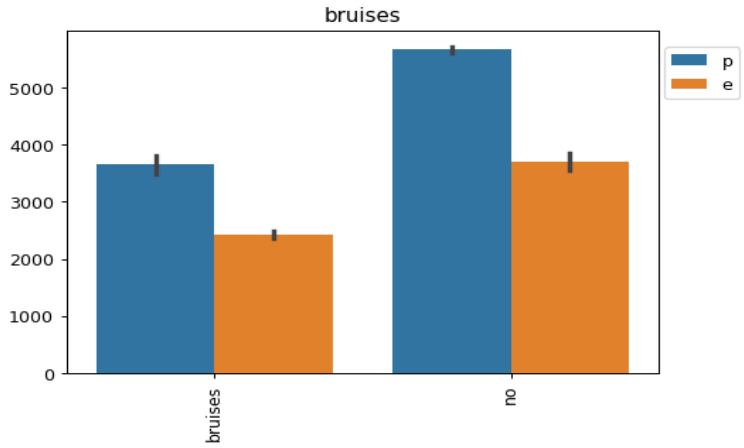
There don't seem to be green and purple cap-colored mushrooms that are poisonous on this dataset. This could be a strong indicator that green colored mushroom caps could result in being edible.



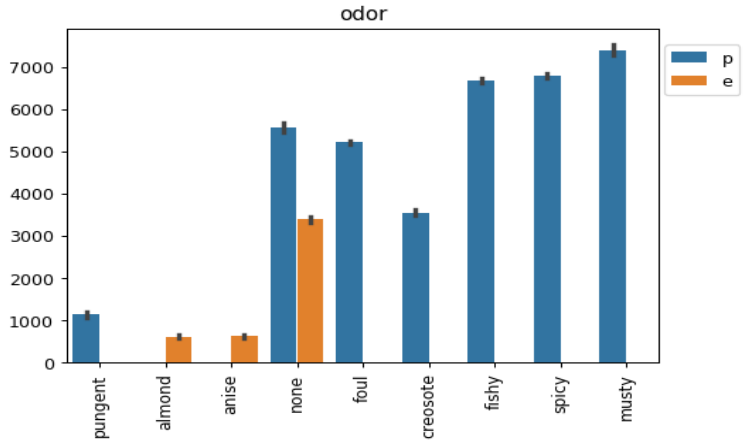
Important note this should be treated with caution. This dataset has a small quantity of information and does not contain all mushroom varieties. The majority of yellow and red mushroom caps are poisonous. Color could be an indication but it should be taken with caution. Some mushrooms change color when picked or bruised due to oxidation.



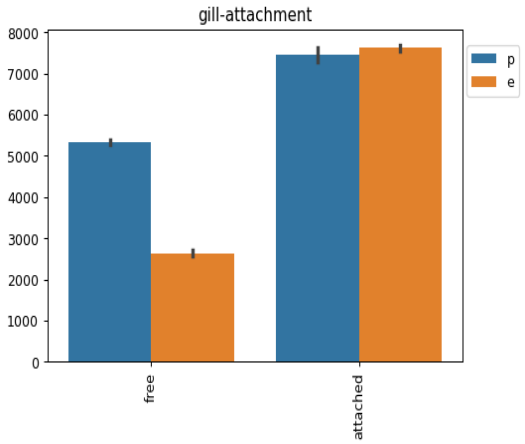
Having bruises appears on both classes and it is higher on the poisonous class. Although the edible class has an appearance on both for at least half of the values contained on the poisonous class.

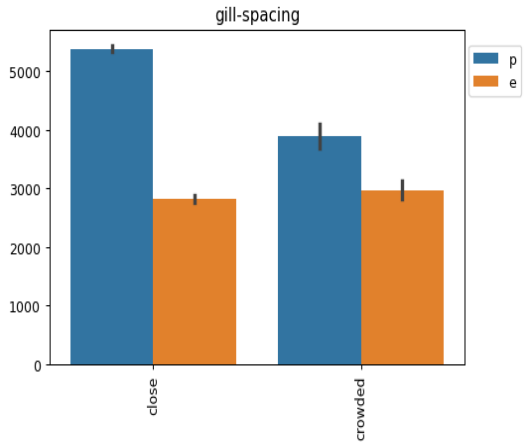


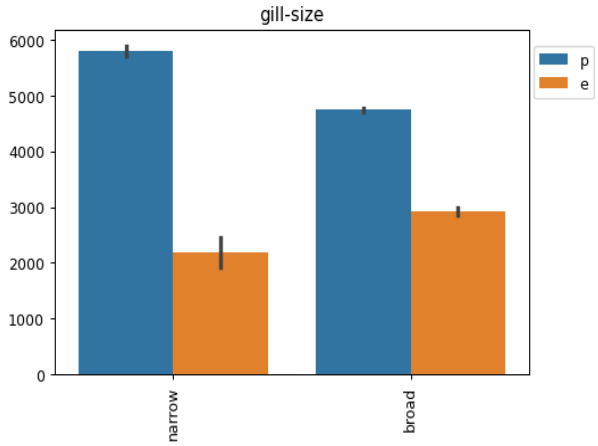
One of the most important distinctions between classes seems to be the odor.
Most poisonous mushrooms seem to have some sort of odor. Especially when it involves a foul, fishy, spicy or musty odor. A few edible mushrooms have an anise or almond odor, but most of them do not have any odor. Both classes have mushrooms also don't have an odor.



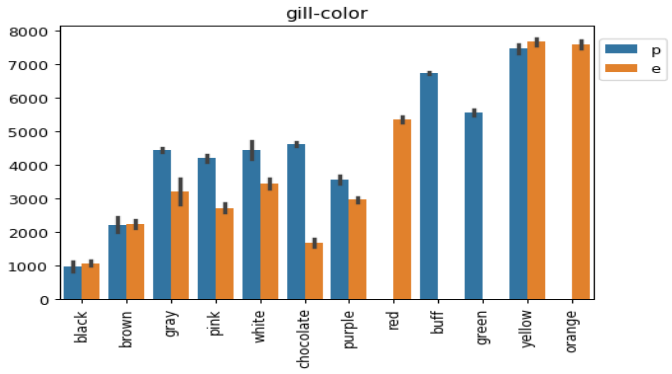
Gill information (attachment, spacing and size) do not seem to be a good indicator of class distinction since both of the classes appear on them.
A close gill spacing and a free gill attachment could be the identifiers for class distinction as well as gill size being narrow or broad.





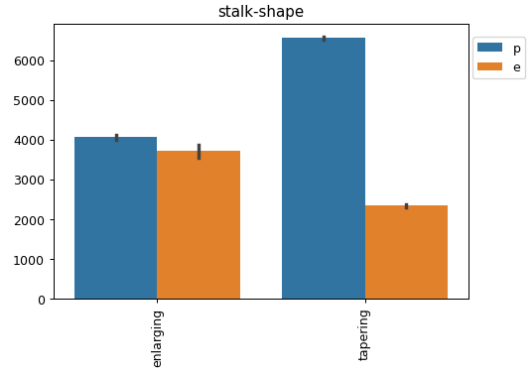


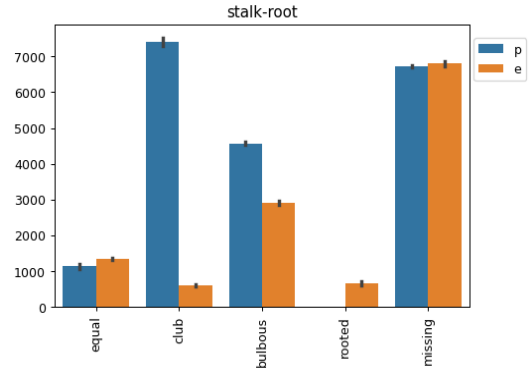
There are several class distinctions related to gill color.
Distinctive feature descriptors for gill color for poisonous is buff and green. For the edible class is orange and red. Chocolate is more prevalent on the poisonous class by more than half of the rows.

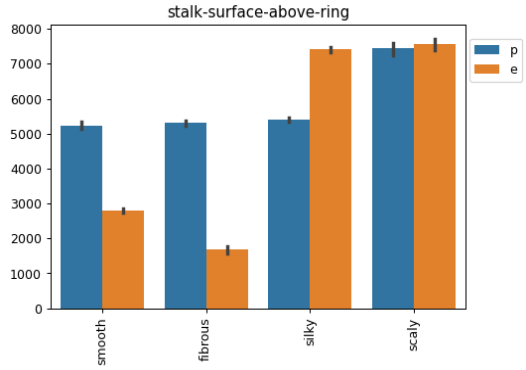


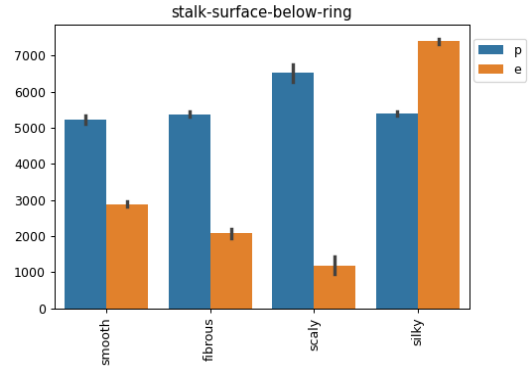
Most of the poisonous mushrooms have:

- A tapered stalk shape.
- Stalk roots that are shaped as a club.
- Stalk surface above ring is usually fibrous. A smooth one is also prevalent but only on almost half of the cases.
- Stalk surface below ring is usually scaly, fibrous or smooth.

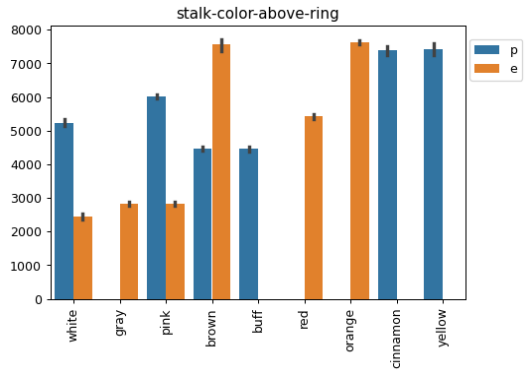


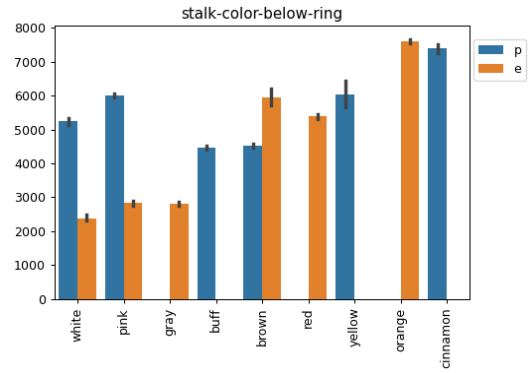




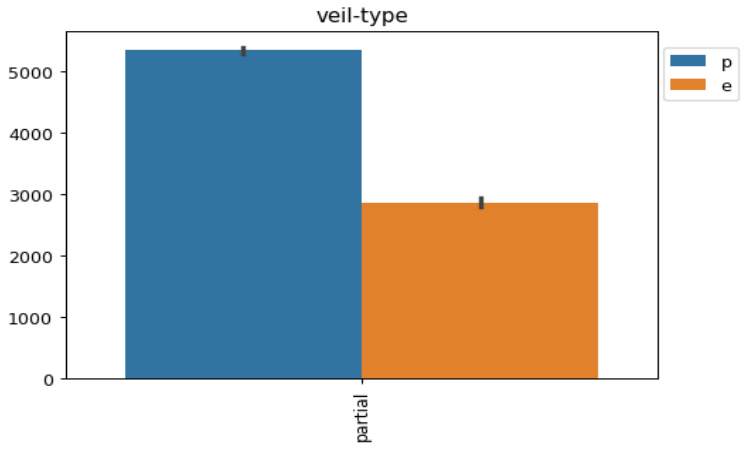


- Stalk color above ring for poisonous is cinnamon or yellow.
- The most common stalk color above ring for edible class is red, orange or brown.
- Stalk color below ring for poisonous is pink or white.
- There does not seem to be a distinctive stalk color below ring for edible class that clearly differentiates both classes.



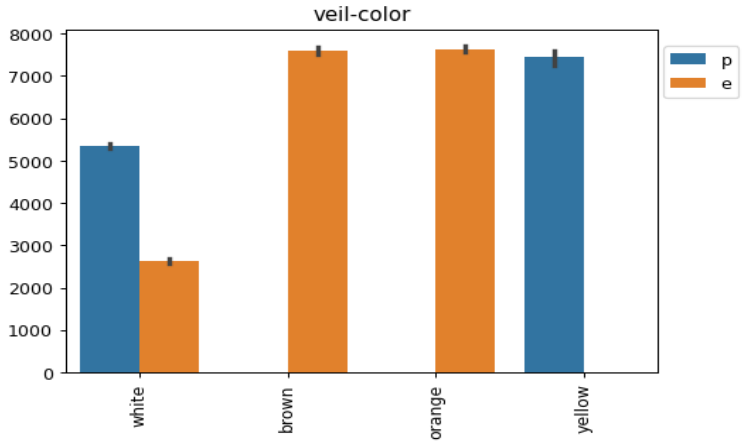


Veil type does not seem to be a good indicator for class identification.

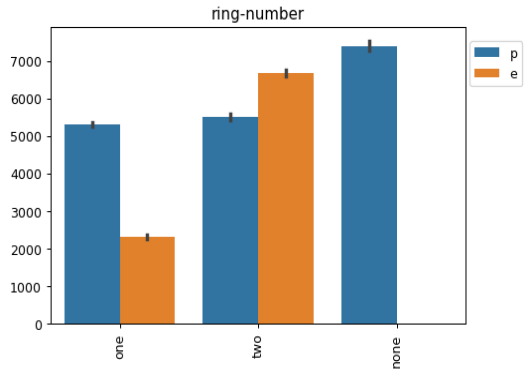


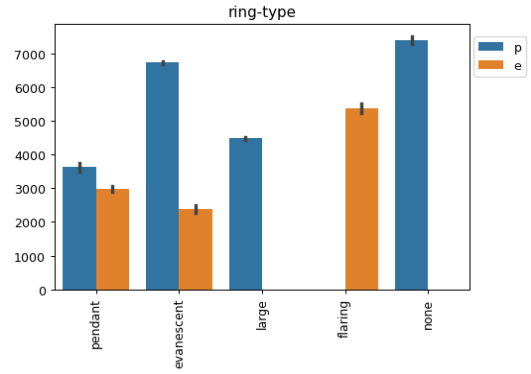
Veil Color

- For poisonous class is predominantly yellow.
- For edible class is brown or orange.

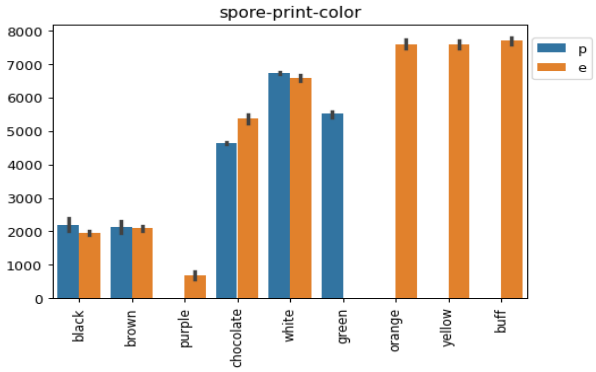


Ring number does not seem to be a good indicator for class distinction.
Poisonous mushrooms do not have a ring and if they do they either have an evanescent or large one.
For edible mushrooms they appear to have a flaring ring type that poisonous do not seem to have.

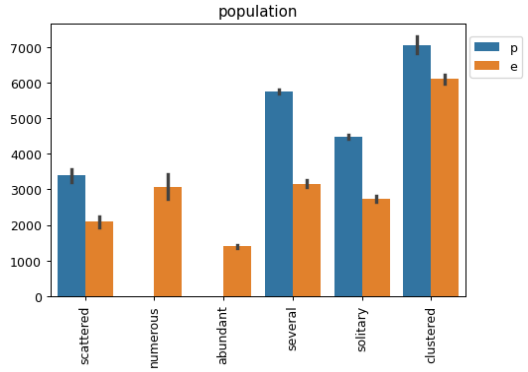


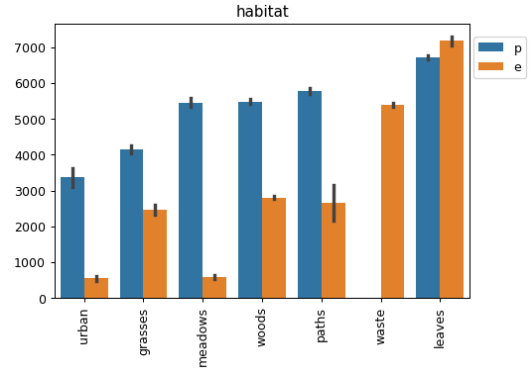


There are class distinctions on the color of the spore prints.
Edible ones have an orange, yellow or buff color while poisonous have a green spore print color.



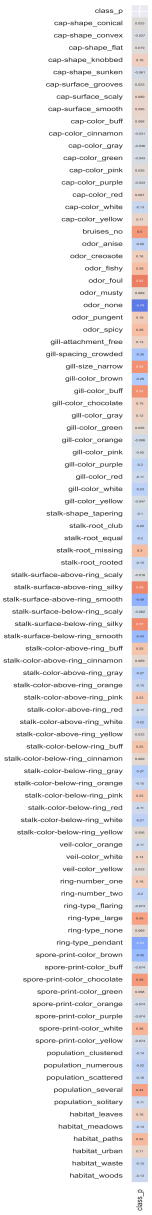
Population does not seem to be the best class differentiator.
"Numerous", "abundand", and "several" are synonyms appearing on the sub feature names.
The habitat in which these mushrooms grow is also confusing.
The class differentiator for the edible class seems to be waste, and for poisonous seems to be meadows.
There might be other factors contributing to habitat that are not being taken into account.





0.2.3 Correlation

Correlation between Class Label Poisonous and other features.



0.3 Objectives

Objectives to meet

1. Create different classification models.
   Elbow method with a range of values.
   Specific hyperparameter selection.
2. Selection of the best performing models.
3. Selection of the most important features (according to the given model).

0.4 Model Creation Procedure

- Data loading.
- Data preprocessing.
  Train Test Split.
- Model creation.
- Performance measurement.
  Hyperparameter tuning.
- Model feature analysis

Data Loading

Data loaded from dataset.

Data preprocessing

Creation of Train Validation Test Split.

Train Test Split	Percentage Split
Train	**80 %**
Validation	**10 %**
Test	**10 %**

0.5 Models

To prevent overfitting on Random Forests:

- Reducing tree depth.
- Reducing the number of variables sampled at each split.
- Using more data.

model_forest_elbow_01

Random Forest Classifier - model_forest_elbow_01 - Range: 1-100
Has an error rate of 0 on validation and test sets.

model_forest_elbow_02

Random Forest Classifier - model_forest_elbow_02 - Range: 1-40
Narrowing number of estimators down from 1 to 40.
Has an error rate of 0 on validation and test sets.
All have an accuracy of 100%, there is no error signal spiking anywhere.

model_forest_elbow_03

Random Forest Classifier - model_forest_elbow_03 - Range: 1-5
Has an error rate of 0 on validation and test sets.

model_forest_04

Random Forest Classifier - model_forest_04 - Specific Parameters

- n_estimators n
- criterion gini
- max_depth None
- min_samples_split 2
- min_samples_leaf 1
- min_weight_fraction_leaf 0
- max_features sqrt
- max_leaf_nodes None
- min_impurity_decrease 0
- bootstrap True
- oob_score False
- n_jobs None
- random_state None
- verbose 0
- warm_start False
- class_weight None
- ccp_alpha 0
- max_samples None

model_forest_05

Random Forest Classifier - model_forest_05 - Specific Parameters

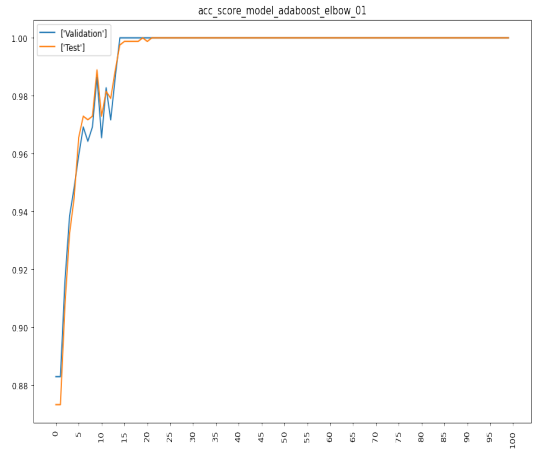
- n_estimators n
- criterion gini
- max_depth None
- min_samples_split 2
- min_samples_leaf 1
- min_weight_fraction_leaf 0
- max_features sqrt
- max_leaf_nodes None
- min_impurity_decrease 0
- bootstrap True
- oob_score False
- n_jobs None
- random_state None
- verbose 0
- warm_start False
- class_weight None
- ccp_alpha 0
- max_samples None

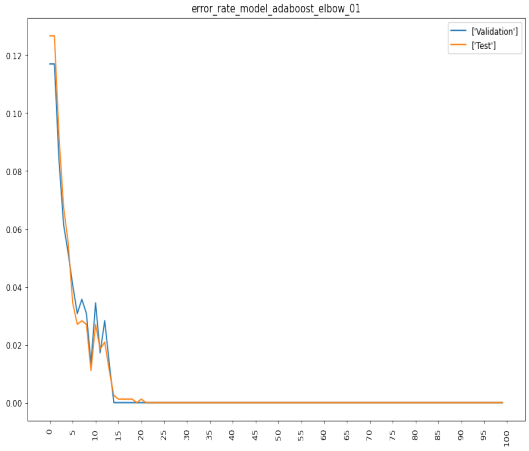
model_adaboost_06

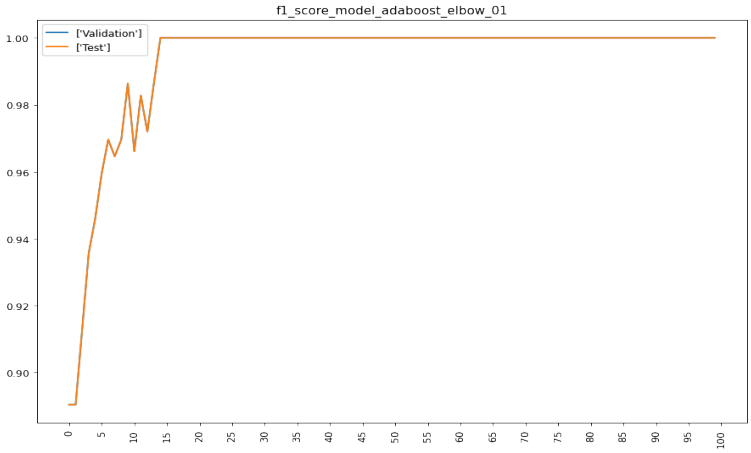
AdaBoost Classifier - model_adaboost_elbow_01 - Range: 1-100

- n_estimators n
- learning_rate 1
- algorithm SAMME.R
- random_state None

Choosing a number of estimators greater than 5 starts to increase accuracy and decrease the error rate. The error rate keeps decreasing from estimator 5 onwards and goes to 0 after estimator 14.





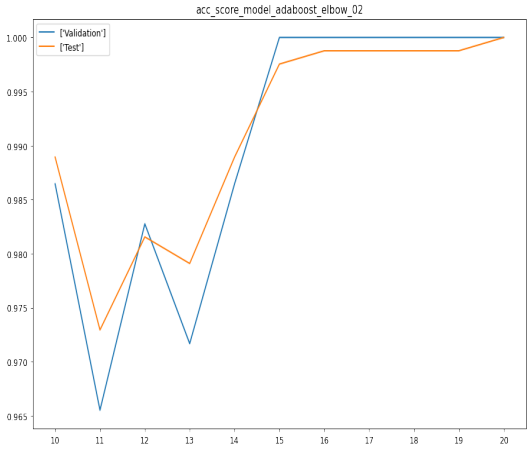


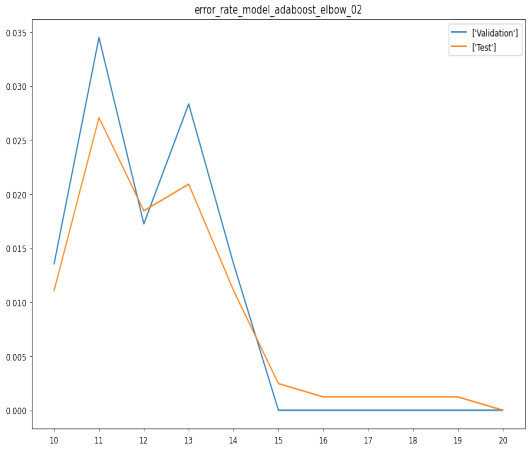
model_adaboost_07

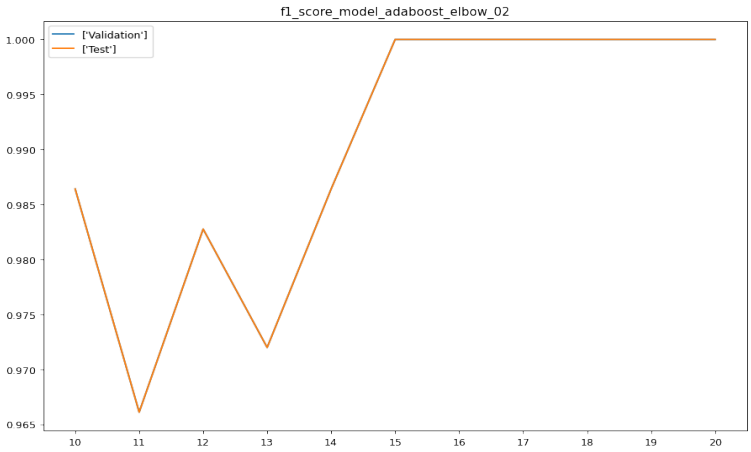
AdaBoost Classifier - model_adaboost_elbow_02 - Range: 10-20

- estimators None
- learning_rate 1
- random_state None
- n_estimators n
- algorithm SAMME.R

There is a spike increase on error from around estimators from 10 and 11 and then again from 12 to 13. The error rate starts decreasing from the estimator number 13 onwards.
The error rate for 15 estimators ranges from 0.004 for test set to 0.001 on the validation set
If a lower number of estimators is required, 12 estimators seems to be a good alternative.







model_adaboost_08

AdaBoost Classifier - model_adaboost_elbow_03 - Specific Parameters

- estimator None
- learning_rate 1
- random_state None
- n_estimators 15
- algorithm SAMME.R

Classification Report Validation Set				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	413
1	1.00	1.00	1.00	399
accuracy			1.00	812
macro avg	1.00	1.00	1.00	812
weighted avg	1.00	1.00	1.00	812

Classification Report Test Set				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	430
1	1.00	1.00	1.00	383
accuracy			1.00	813
macro avg	1.00	1.00	1.00	813
weighted avg	1.00	1.00	1.00	813

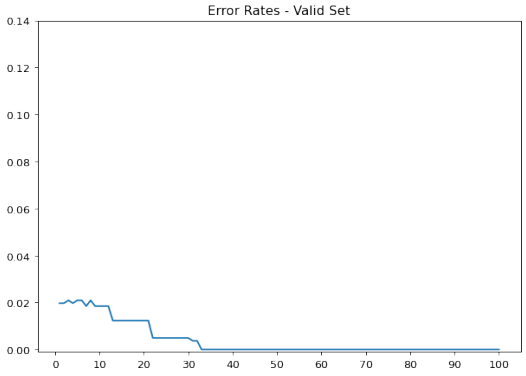
model_gradboost_09

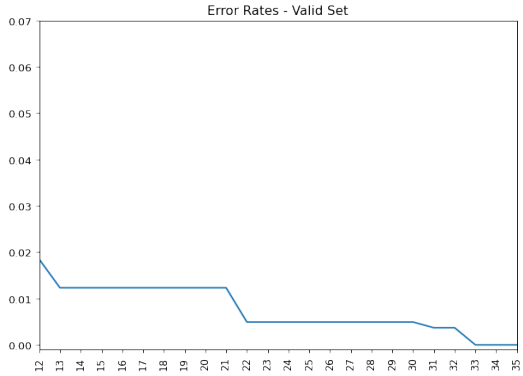
Gradient Boosting Classifier - model_gradboost_elbow_01 - Range: 1-101

- loss log_loss
- learning_rate 0.1
- n_estimators n
- subsample 1.0
- criterion friedman_mse
- min_samples_split 2
- min_samples_leaf 1
- min_weight_fraction_leaf 0
- max_depth 3
- min_impurity_decrease 0.0
- init None
- random_state None
- max_features None
- verbose 0
- max_leaf_nodes None
- warm_start False
- validation_fraction 0.1
- n_iter_no_change None
- tol 0.0001
- ccp_alpha 0.0

Error Rate for Validation Set

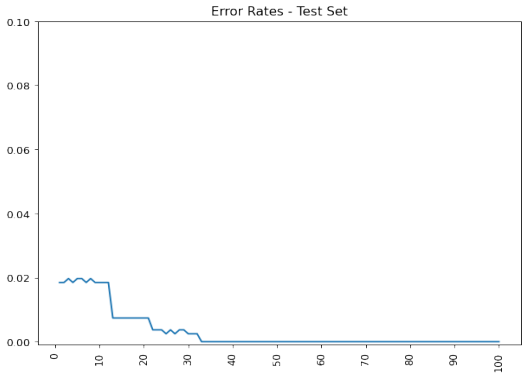
The error is low in all estimators (Around 0.02), but keeps decreasing from estimator number 13 onwards.

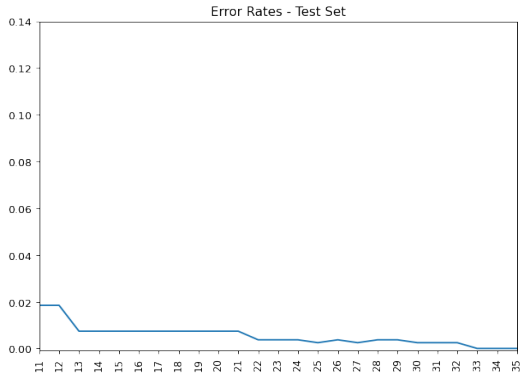




Error Rate for Test Set

The error is low in all estimators (Around 0.02), but keeps decreasing from estimator number 12 onwards.





0.5.1 Models Summary

Models were created. First using the elbow method to find the range of best performing parameters. Several models were created narrowing down a specific set of parameters according to previous models or information.
All Models (Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier) created performed well on the metrics reported.

0.5.2 Model Selection

The following models will be used for feature importance selection:

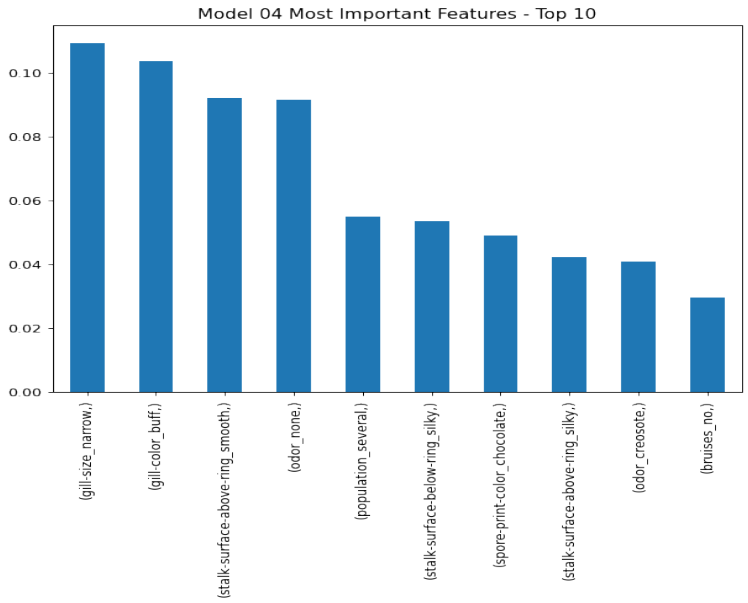
- Random Forest Classifier - model_forest_04 - Specific Parameters
- Random Forest Classifier - model_forest_05 - Specific Parameters

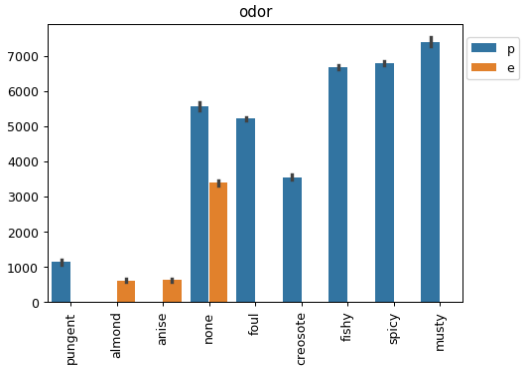
0.6 Feature Importance

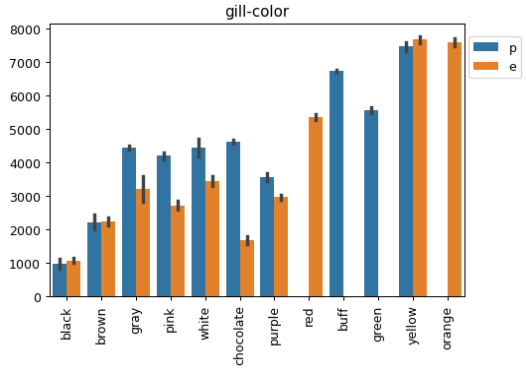
Feature Importance for Model 04

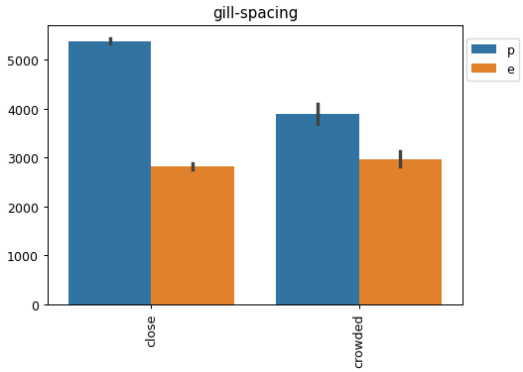
Random Forest Classifier - model_forest_04 - Specific Parameters
Filtering the model selection importance by features that have more than 1% of importance.

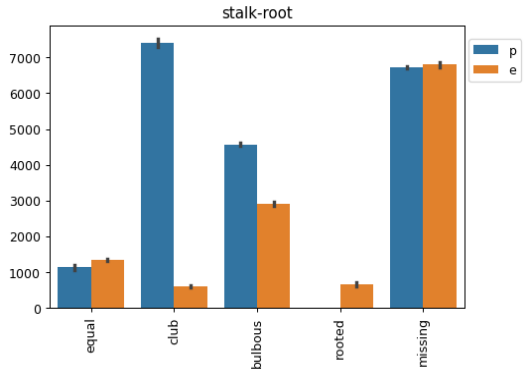
No.	Feature	Percentage	Sum Percentage
1	odor_none	0.292642	10.94
2	odor_foul	0.110902	21.29
3	gill-color_buff	0.089641	30.51
4	spore-print-color_chocolate	0.073211	39.67
5	gill-spacing_crowded	0.062453	45.17
6	stalk-root_club	0.061218	50.51
7	gill-size_narrow	0.061010	55.39
8	bruises_no	0.055607	59.61
9	odor_creosote	0.019430	63.68
10	odor_pungent	0.018778	66.63

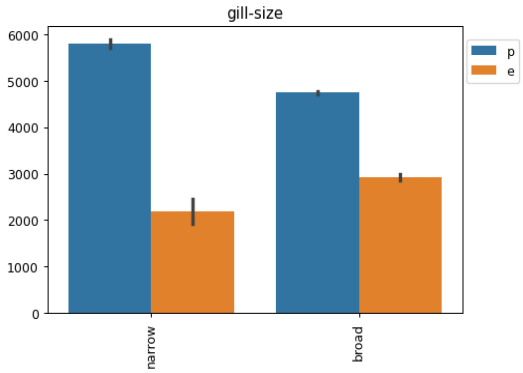


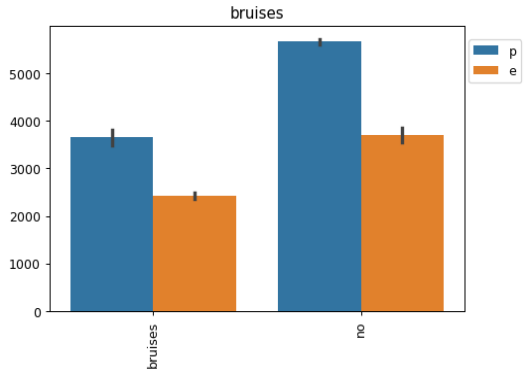

odor
p
e
0
1000
2000
3000
4000
5000
6000
7000
pungent
almond
anise
none
foul
creosote
fishy
spicy
musty


gill-color
p
e
0
1000
2000
3000
4000
5000
6000
7000
8000
black
brown
gray
pink
white
chocolate
purple
red
buff
green
yellow
orange


gill-spacing
p
e
0
1000
2000
3000
4000
5000
close
crowded


stalk-root
p
e
0
1000
2000
3000
4000
5000
6000
7000
equal
club
bulbous
rooted
missing

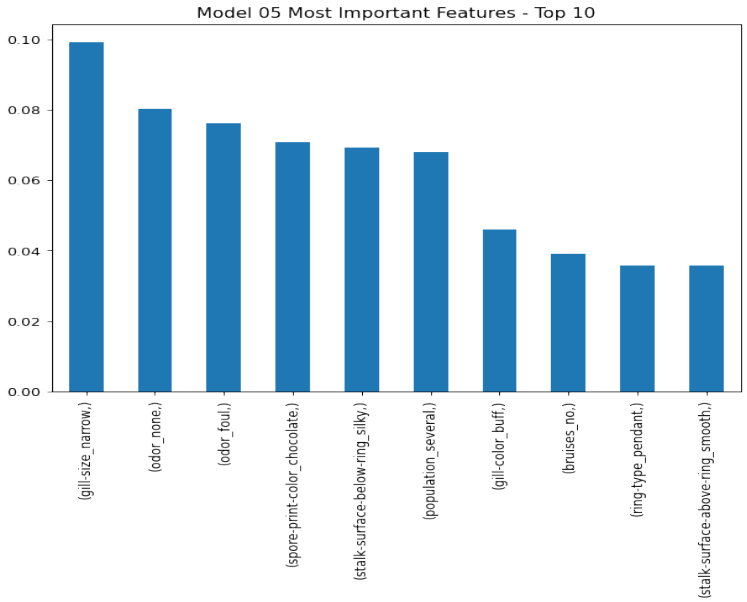

gill-size
p
e
0
1000
2000
3000
4000
5000
6000
narrow
broad

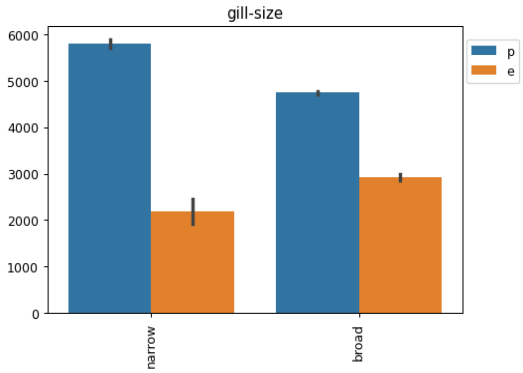

bruises
p
e
0
1000
2000
3000
4000
5000
bruises
no

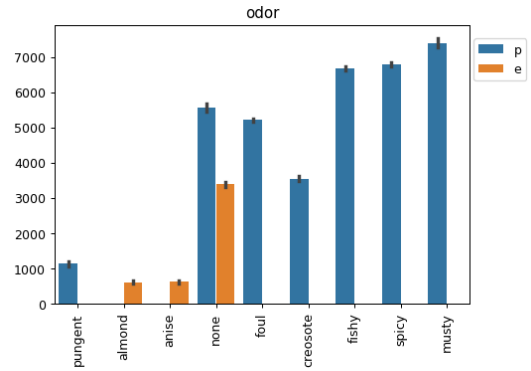
Feature Importance for Model 05

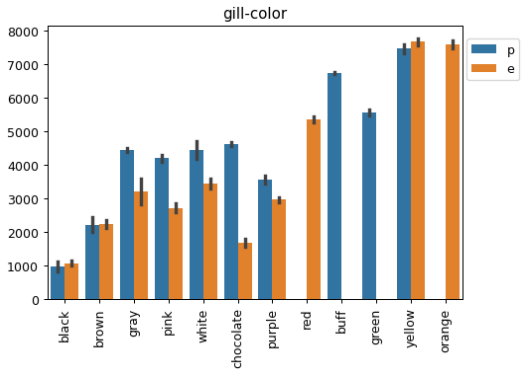
Random Forest Classifier - model_forest_05 - Specific Parameters
Filtering the model selection importance by features that have more than 1% of importance.

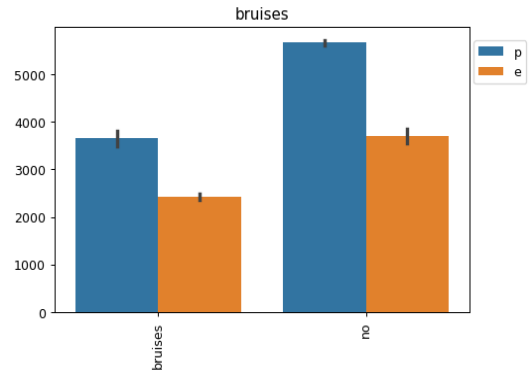
No.	Feature	Percentage	Sum Percentage
1	gill-size_narrow	0.127468	9.93
2	odor_foul	0.092675	17.96
3	gill-color_buff	0.069015	25.57
4	bruises_no	0.056847	32.64
5	odor_none	0.050715	39.56
6	stalk-surface-below-ring_silky	0.046044	46.36
7	stalk-shape_tapering	0.043962	50.96
8	gill-spacing_crowded	0.043039	54.87
9	ring-type_pendant	0.042135	58.45
10	stalk-surface-above-ring_smooth	0.039530	62.01

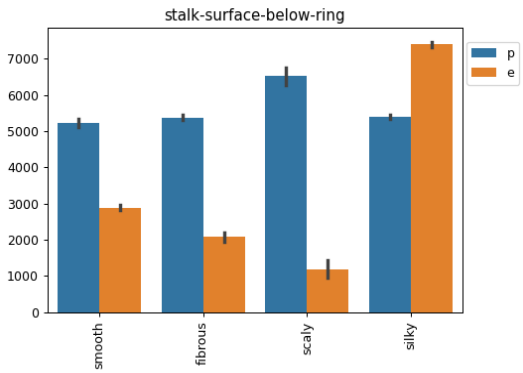


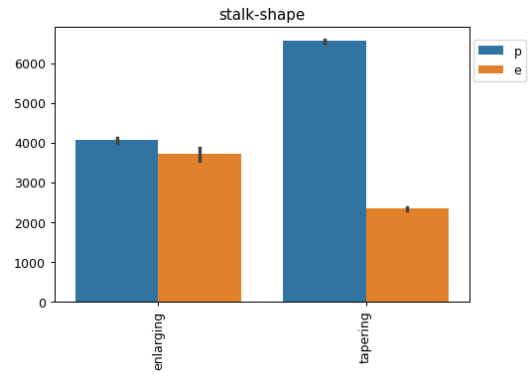

gill-size
p
e
narrow
broad

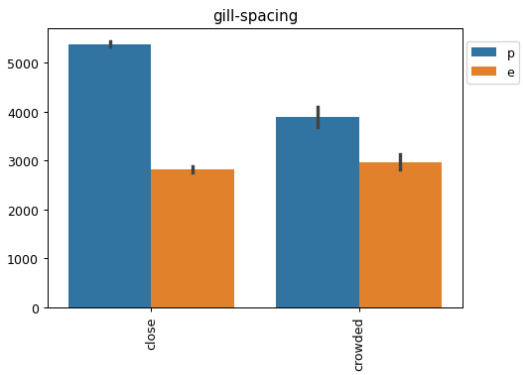

odor
p
e
pungent
almond
anise
none
foul
creosote
fishy
spicy
musty

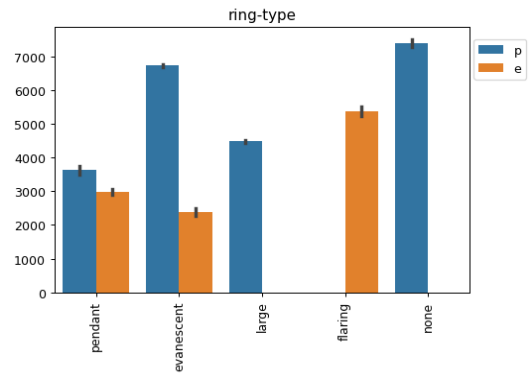

gill-color
p
e
black
brown
gray
pink
white
chocolate
purple
red
buff
green
yellow
orange

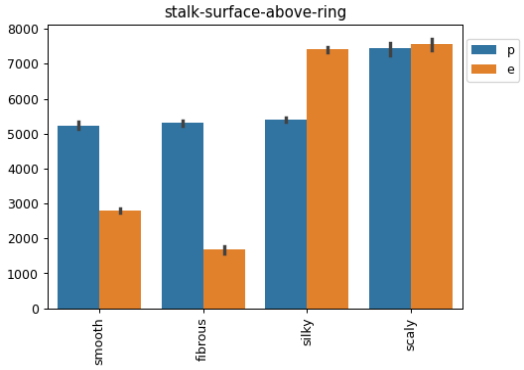

bruises
p
e
bruises
no


stalk-surface-below-ring
p
e
smooth
fibrous
scaly
silky


stalk-shape
p
e
enlarging
tapering


gill-spacing
p
e
close
crowded

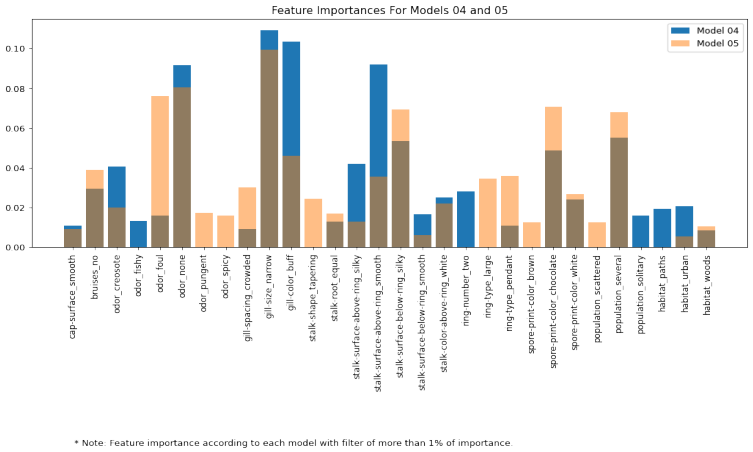

ring-type
p
e
pendant
evanescent
large
flaring
none



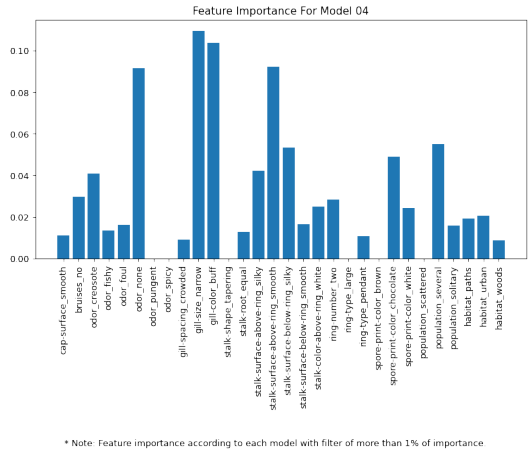
0.7 Feature Selection Summary

Selection filter of features that have an importance greater than 1% to then select the ones that have the most weight. The decision to choose 10 of the most important features was arbitrary and other number could be selected. The top 10 most important features cover more than 60% of the importance for each model.

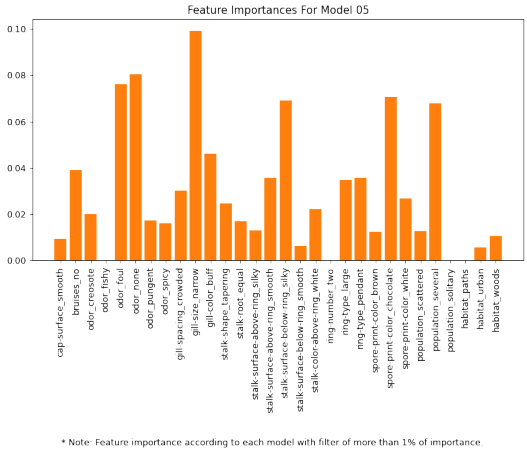
The top 10 features selected for model 04 cover around 66% model 05 cover around 62% of what each of the models consider important for a class label identification.



* Note: Feature importance according to each model with filter of more than 1% of importance.

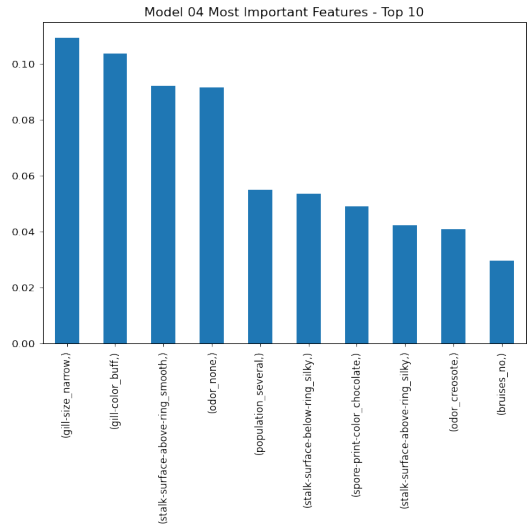


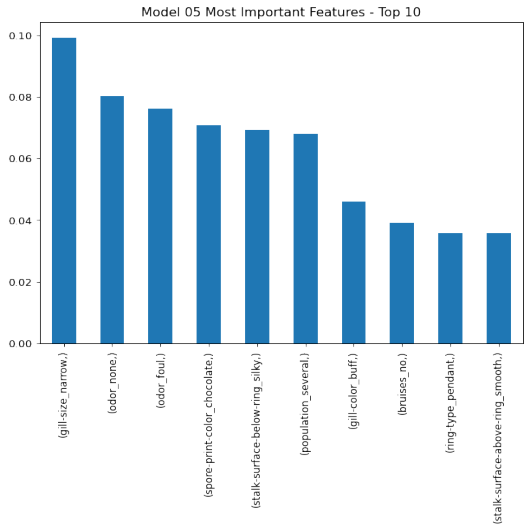
* Note: Feature importance according to each model with filter of more than 1% of importance



* Note: Feature importance according to each model with filter of more than 1% of importance

Feature Importance For Selected Models Sorted





0.8 Conclusion

According to models selected, the most important features to identify whether a mushroom belongs to the poisonous or edible class are odor, gill color, spore print color, gill spacing, whether it has bruises or not. Stalk shape and stalk surface (below and above the ring) also play an important role.

Both models took more importance on some of the features that on the first analysis were thought out to be less important. An example of this was bruises. The poisonous mushrooms has more values on both of these subfeatures than the edible ones. The edible class still has at least half of the values on both of those subfeatures.

Some other features were thought to have a substantial impact from the first analysis. Odor is a clear example of this. There are clear distinctions between the poisonous and edible classes on a lot of subfeatures.

Having an odor, especially a foul, creosote, musty, fishy, spicy clearly distinguishes a poisonous versus an edible, according to the information provided.

Choosing the top 10 most important features was arbitrary. It was used to select as few features as possible that could cover more than a 50% of importance.

Both models have more than 50% of importance covered with the selection of the top ten features. Model 04 covers around around 66%, and model 05 covers around 62% of the percentage of importance of all the features included on this dataset.

The top 10 features selected for model 04 cover around 66% model 05 cover around 62% of what each of the models consider important for a class label identification.

The decision to choose 10 features was to select the most important ones that can outweight a guessing of 50% 50%.

This with the intent of using a fast selection process on the beginning, to then rigoruously detect whether a mushroom is edible on the second inspection.

Both models have more than 50% of importance covered with the selection of the top ten features. The more features used to distinguish a poisonous versus an edible mushroom will always be better, since this will filter out, especially the ones that have a similar look or have particular similar features. This commonly happens amongst several groups of mushrooms.

The more rigorous and the more information the user has on the inspection, the more certainty when identifying an edible against a poisonous one.

0.8.1 Side Notes

It is likely that other factors not contained on the datasets have an important effect on whether it is from one class or the other. Some features found on the dataset were thought to be confusing. Especially on the "population" feature. "numerous", "several" and "abundant" seem to have a similar meaning. This could affect the decision making process for the algorithm.

Some other information that might affect the color of the mushroom is the mushroom oxidation. Some mushrooms change color after being picked due to the oxidative process.

References

Mushroom (1987), UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5959T.