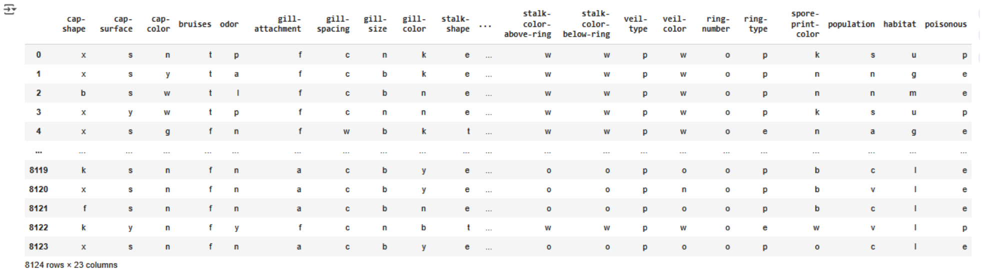
DS 5101 - Computer Programming for Data Science and AI

Assignment

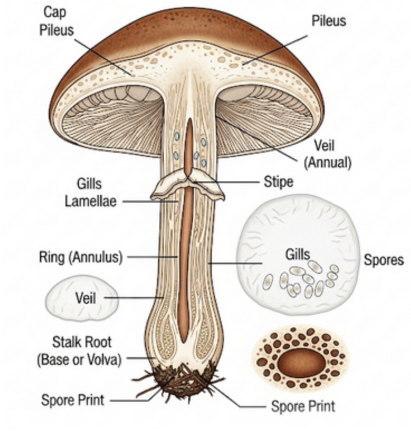
Initial Exploration

Task	Building classification model using the ‘Mushroom’ dataset from the UCI repository
Dataset	https://archive.ics.uci.edu/dataset/73/mushroom
Dataset Characteristics	Multivariate with 22 features
Subject Area	Biology
Number of instances	8124
Class Labels	edible=e, poisonous=p
Feature Type	Categorical
Dataset	

	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	...	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population	habitat	poisonous
0	x	s	n	t	p	f	c	n	k	e	...	w	w	p	w	o	p	k	s	u	p
1	x	s	y	t	a	f	c	b	k	e	...	w	w	p	w	o	p	n	n	g	e
2	b	s	w	t	l	f	c	b	n	e	...	w	w	p	w	o	p	n	n	m	e
3	x	y	w	t	p	f	c	n	n	e	...	w	w	p	w	o	p	k	s	u	p
4	x	s	g	f	n	f	w	b	k	t	...	w	w	p	w	o	e	n	a	g	e
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8119	k	s	n	f	n	a	c	b	y	e	...	o	o	p	o	o	p	b	c	l	e
8120	x	s	n	f	n	a	c	b	y	e	...	o	o	p	n	o	p	b	v	l	e
8121	f	s	n	f	n	a	c	b	n	e	...	o	o	p	o	o	p	b	c	l	e
8122	k	y	n	f	y	f	c	n	b	t	...	w	w	p	w	o	e	w	v	l	p
8123	x	s	n	f	n	a	c	b	y	e	...	o	o	p	o	o	p	o	c	l	e

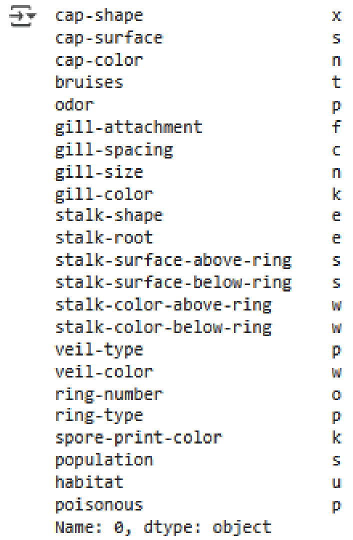
8124 rows × 23 columns

Anatomy



Data interpretation

For example, let’s consider the first row of the dataset

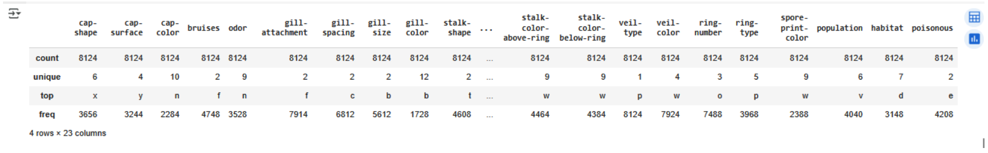


```
cap-shape                   x
cap-surface                 s
cap-color                   n
bruises                     t
odor                        p
gill-attachment             f
gill-spacing                c
gill-size                   n
gill-color                  k
stalk-shape                 e
stalk-root                  e
stalk-surface-above-ring    s
stalk-surface-below-ring    s
stalk-color-above-ring      w
stalk-color-below-ring      w
veil-type                   p
veil-color                  w
ring-number                 o
ring-type                   p
spore-print-color           k
population                  s
habitat                     u
poisonous                   p
Name: 0, dtype: object
```

Here is the interpretation,

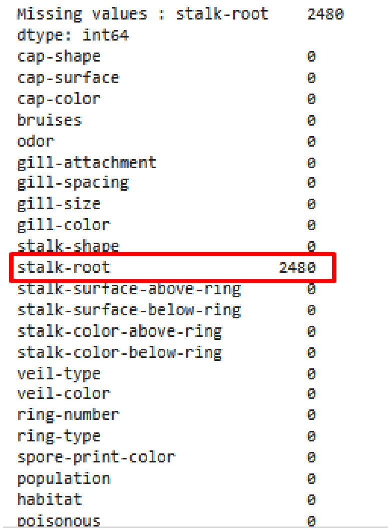
This mushroom has a convex (x), smooth (s), brown (n) cap with bruises (t) and a pungent odor (p). Its gills are freely attached (f), closely spaced (c), narrow (n), and black (k). The stalk enlarges (e) with an equal base (e), smooth surfaces above and below the ring (s,s), white colors on both sides (w,w), a partial veil (p) that's white (w), one ring (o) of the pendant type (p), black spore prints (k), grows in scattered groups (s) in urban areas (u), and is poisonous (p).

Statistic summery



	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	stalk-shape	...	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population	habitat	poisonous
count	8124	8124	8124	8124	8124	8124	8124	8124	8124	8124	...	8124	8124	8124	8124	8124	8124	8124	8124	8124	8124
unique	6	4	10	2	9	2	2	2	12	2	...	9	9	1	4	3	5	9	6	7	2
top	x	y	n	f	n	f	c	b	b	t	...	w	w	p	w	o	p	w	v	d	e
freq	3656	3244	2284	4748	3528	7914	6812	5612	1728	4608	...	4464	4384	8124	7924	7488	3968	2388	4040	3148	4208

4 rows × 23 columns

Missing values	There are 2480 missing values for the column stalk-root 

```
Missing values : stalk-root    2480
dtype: int64
cap-shape                         0
cap-surface                       0
cap-color                         0
bruises                           0
odor                              0
gill-attachment                   0
gill-spacing                      0
gill-size                         0
gill-color                        0
stalk-shape                       0
stalk-root                     2480
stalk-surface-above-ring          0
stalk-surface-below-ring          0
stalk-color-above-ring            0
stalk-color-below-ring            0
veil-type                         0
veil-color                        0
ring-number                       0
ring-type                         0
spore-print-color                 0
population                        0
habitat                           0
poisonous                         0
```

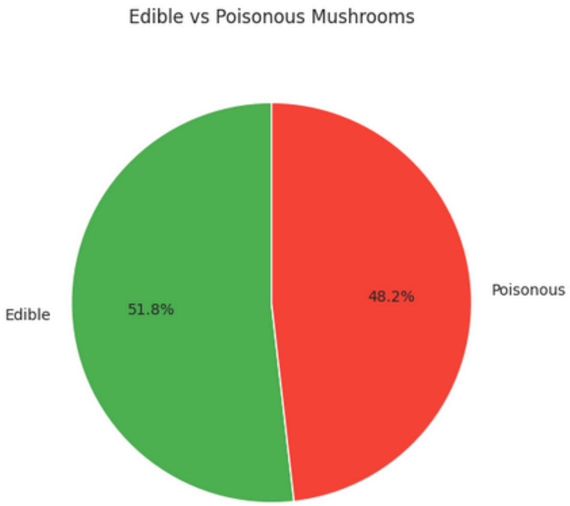
Data columns and value codes

Feature Name	**Codes**	**Meanings**
cap-shape	b,c,x,f,k,s	bell, conical, convex, flat, knobbed, sunken
cap-surface	f,g,y,s	fibrous, grooves, scaly, smooth
cap-color	n,b,c,g,r,p,u,e,w,y	brown, buff, cinnamon, gray, green, pink, purple, red, white, yellow
bruises?	t,f	bruises present, no bruises
odor	a,l,c,y,f,m,n,p,s	almond, anise, creosote, fishy, foul, musty, none, pungent, spicy
gill-attachment	a,d,f,n	attached, descending, free, notched
gill-spacing	c,w,d	close, crowded, distant
gill-size	b,n	broad, narrow
gill-color	k,n,b,h,g,r,o,p,u,e,w,y	black, brown, buff, chocolate, gray, green, orange, pink, purple, red, white, yellow
stalk-shape	e,t	enlarging, tapering
stalk-root	b,c,u,e,z,r,?	bulbous, club, cup, equal, rhizomorphs, rooted, missing

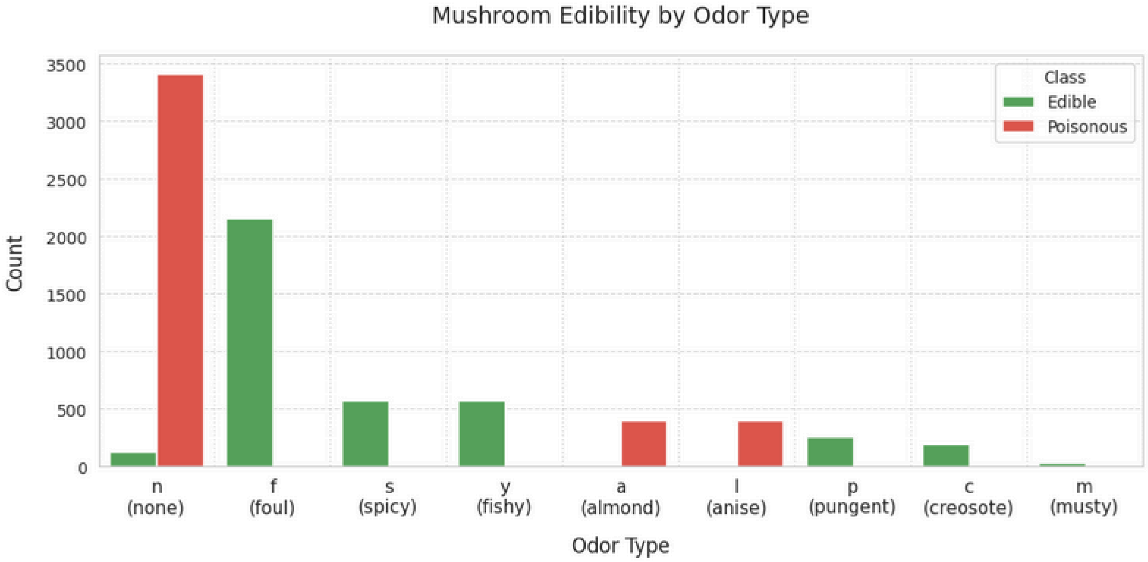
stalk-surface-above-ring/ stalk-surface-below-ring	f,y,k,s	fibrous, scaly, silky, smooth
stalk-color-above-ring/ stalk-color-below-ring	n,b,c,g,o,p,e,w,y	brown, buff, cinnamon, gray, orange, pink, red, white, yellow
veil-type	p,u	partial, universal
veil-color	n,o,w,y	brown, orange, white, yellow
ring-number	n,o,t	none, one, two
ring-type	c,e,f,l,n,p,s,z	cobwebby, evanescent, flaring, large, none, pendant, sheathing, zone
spore-print-color	k,n,b,h,r,o,u,w,y	black, brown, buff, chocolate, green, orange, purple, white, yellow
population	a,c,n,s,v,y	abundant, clustered, numerous, scattered, several, solitary
habitat	g,l,m,p,u,w,d	grasses, leaves, meadows, paths, urban, waste, woods
poisonous	**e,p**	**edible, poisonous**

Visualization

When the ‘Poisonous’ column is considered for the plotting, there are 51.8% of edible mushroom samples and 48.2% as poisonous mushrooms. Therefore it can be seen that the sample space is balanced with these two categories.

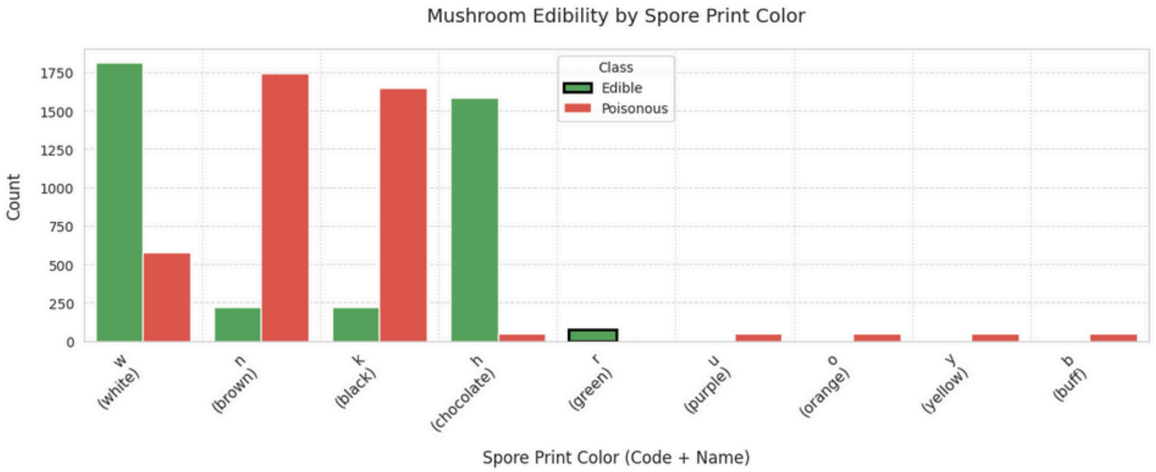


This bar chart titled "Mushroom Edibility by Odor Type" illustrates the distribution of edible and poisonous mushrooms across various odor types. The most striking feature is the dominance of the 'n' (none) odor type, where a significantly higher number of mushrooms are edible compared to poisonous. Conversely, the 'f' (foul) odor type shows a large number of poisonous mushrooms and a relatively small number of edible ones. Other odor types like 's' (spicy), 'y' (fishy), 'p' (pungent), and 'c' (creosote) primarily contain edible mushrooms, while 'a' (almond) and 'l' (anise) are exclusively associated with poisonous mushrooms. The 'm' (musty) odor type has very few mushrooms, all of which are poisonous. This visualization clearly highlights that odor is a strong indicator of a mushroom's edibility, with certain odors being almost exclusively associated with either edible or poisonous varieties.

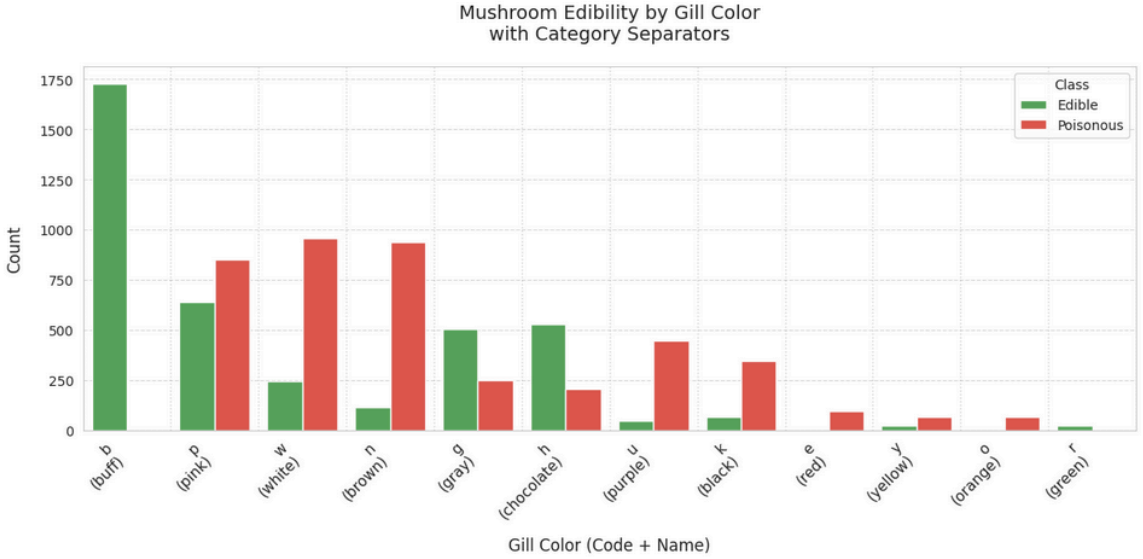


The "Mushroom Edibility by Spore Print Color" chart reveals 'white' spore prints are mostly edible. Conversely, 'brown' and 'black' spore prints are strongly linked to poisonous mushrooms.

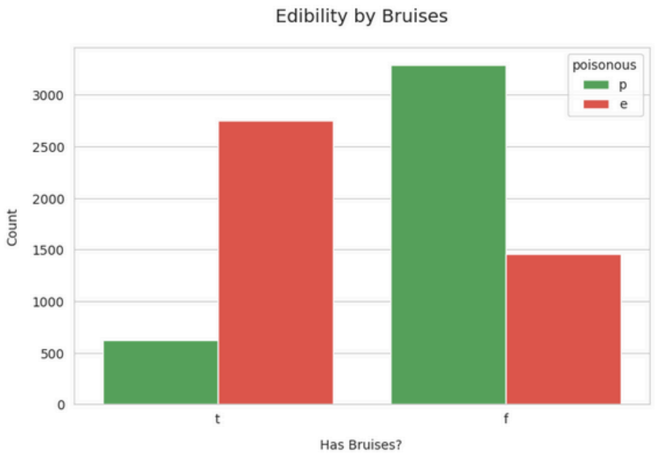
'Chocolate' spore prints indicate many edible mushrooms. Other, less common spore print colors generally point to poisonous varieties, highlighting spore print color as a strong indicator of edibility.



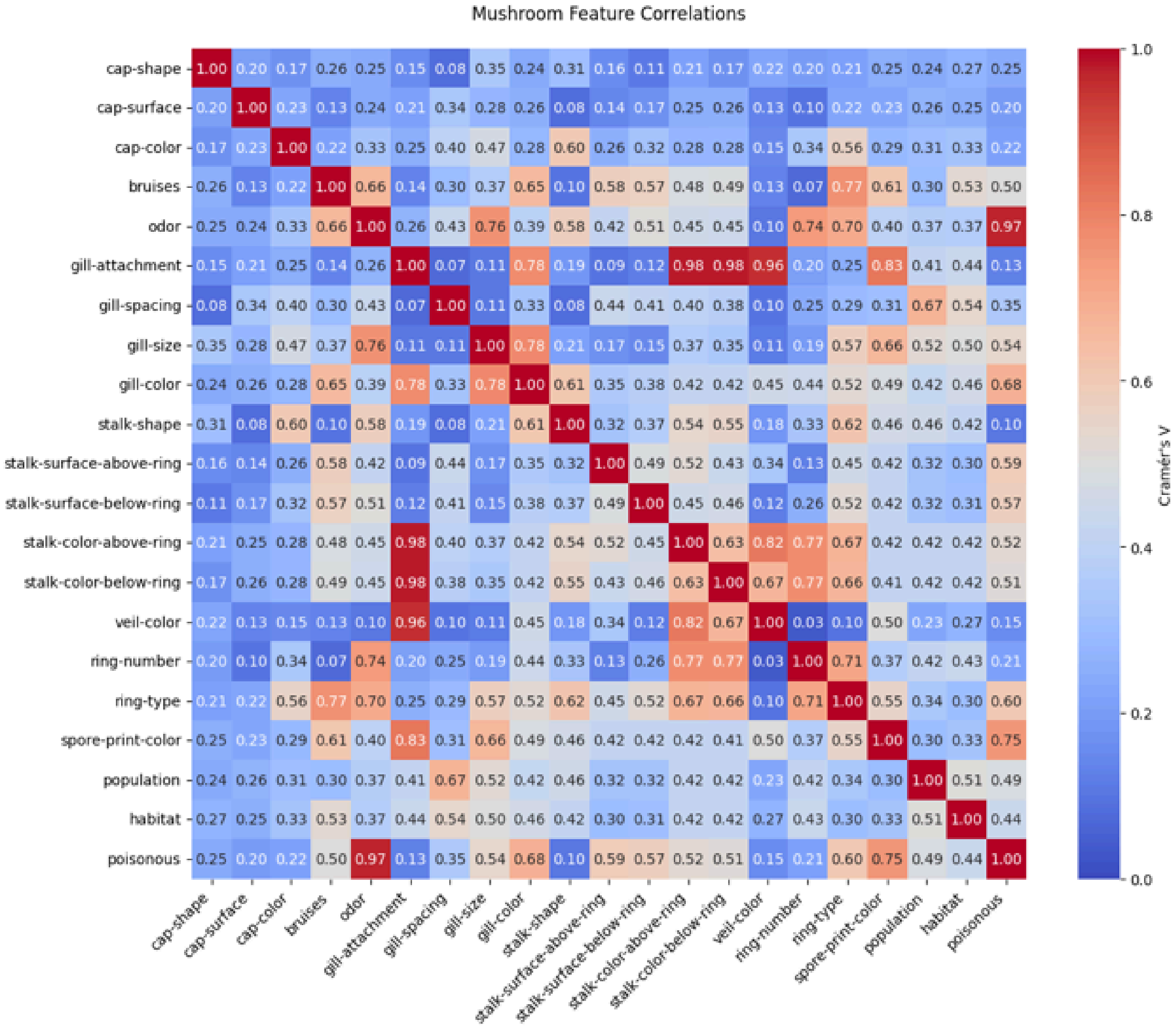
The bar chart on mushroom edibility by gill color shows that 'buff' gills are strongly associated with edible mushrooms, while 'pink,' 'white,' and 'brown' gills are largely linked to poisonous ones. Other colors like 'gray' and 'chocolate' have more edible mushrooms, but less common colors generally indicate poisonous varieties. This suggests gill color is a key factor in mushroom edibility.



The "Edibility by Bruises" chart shows a clear relationship: mushrooms that bruise are more likely to be poisonous, while those that don't bruise tend to be edible. Specifically, a large majority of bruising mushrooms are poisonous. Conversely, non-bruising mushrooms have a significantly higher count of edible varieties, though poisonous ones are still present. This indicates bruising is a strong, but not definitive, indicator of a mushroom's edibility.



This Cramér's V heatmap shows the strength of association between pairs of categorical features in the mushroom dataset. One clear observation is that some features (like odor, gill-size, spore-print-color, and gill-color) exhibit a strong relationship with the target variable "poisonous," as indicated by their high correlation values. These features are likely to be highly predictive in a classification model. In contrast, features like cap-shape, veil-color, and habitat show relatively weak associations with the poison target, suggesting they may contribute less to prediction and could potentially be dropped or deprioritized depending on the modeling goal. (Note that Cramér's V captures only pairwise linear associations and does not account for interactions or multivariate effects.)



Decision tree modeling

According to the above analysis, it was observed that some features in the dataset are directly interconnected with the poisonous target variable. When working with the decision tree model, it should highlight these important features, and trees will be built based on them. The initial categorical data are not accepted by scikit-learn trees; therefore, the encoded (One-Hot Encoding) data will be used to build the tree.

Also, decision trees are quite robust and do not require prior feature selection based on correlation (from the heat map above). Usually, they can automatically perform feature selection during training and data splitting based on information gain (entropy or gini index). Then they can handle irrelevant features gracefully by simply not using them in the tree. However, data investigation and handling missing values are required before modeling the decision tree.

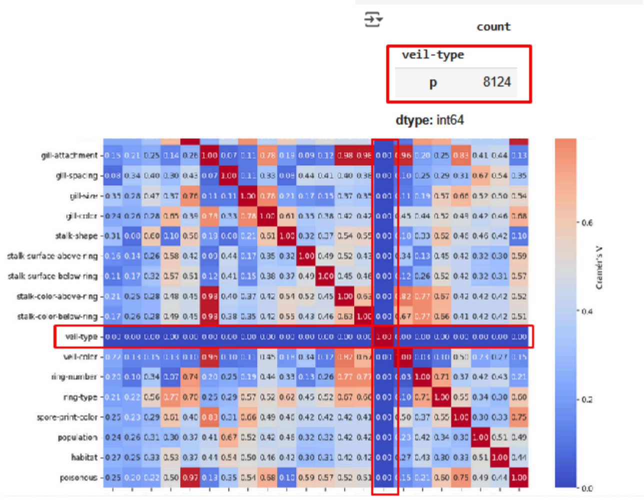
Handling categorical data and missing values.

According to the initial investigation, there are missing values indicated by '?' for the 'stalk-root' feature. It can be seen that there are 2,480 missing values out of 8,124 instances. Since this represents 30% of samples, it is not feasible to drop the related rows. Instead, the entire 'stalk-root' column will be dropped from the initial dataframe to continue with data modeling.

Also, we can see that the entire dataset is categorical; therefore, the dataset will be encoded to get the numerical representation needed to build the data model.

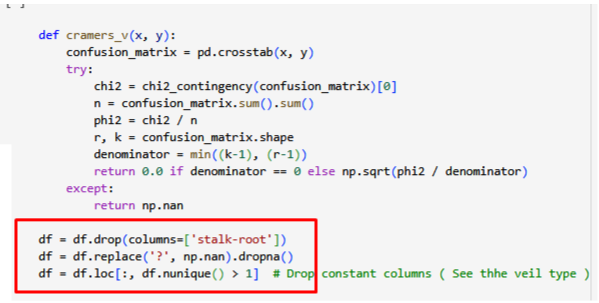
Here, the 'One-Hot Encoding' technique was used to convert the categorical data into binary numerical format that machine learning models can process. With this technique, each categorical value becomes a new binary column, and the initial dimension of the dataframe will be increased. Note that the target is already in binary representation; therefore, that column was not encoded.

When building the heatmap using the 'Cramer's V' technique, it was observed that the heatmap was not built correctly for the 'veil-type' feature. This was because there is only one type of category, which is 'p-partial'.



```
          count
veil-type
p          8124

dtype: int64
```

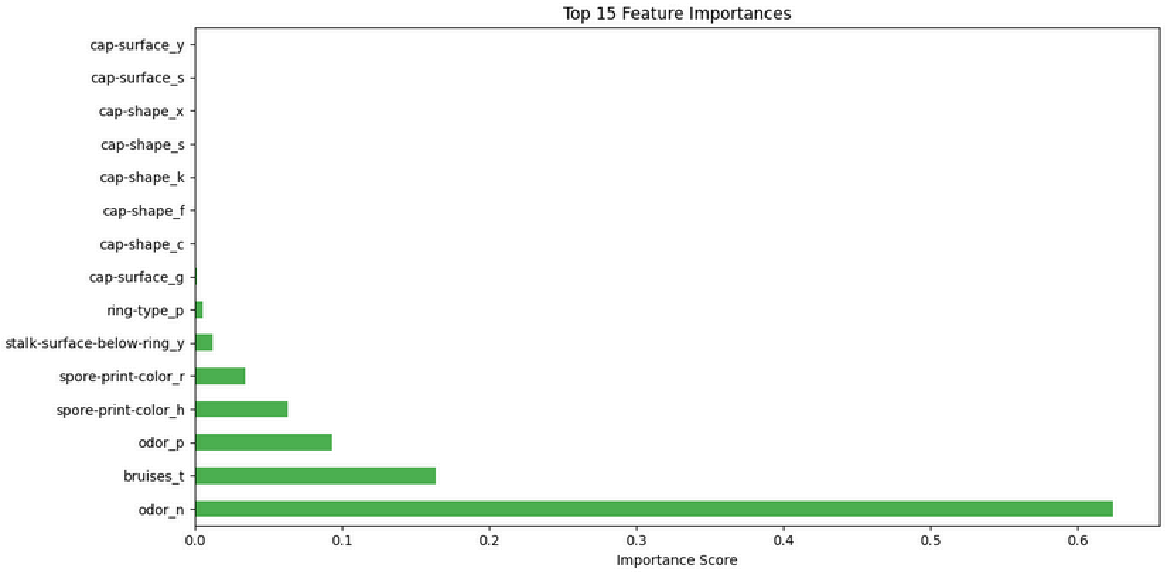


```
def cramers_v(x, y):
    confusion_matrix = pd.crosstab(x, y)
    try:
        chi2 = chi2_contingency(confusion_matrix)[0]
        n = confusion_matrix.sum().sum()
        phi2 = chi2 / n
        r, k = confusion_matrix.shape
        denominator = min((k-1), (r-1))
        return 0.0 if denominator == 0 else np.sqrt(phi2 / denominator)
    except:
        return np.nan

df = df.drop(columns=['stalk-root'])
df = df.replace('?', np.nan).dropna()
df = df.loc[:, df.nunique() > 1]  # Drop constant columns ( See thhe veil type )
```

Therefore, the decision tree modeling excluded both the stalk-root column, which has considerable missing values, and the veil-type column, which contains only 'p-partial' type data.

The initial decision tree model was built using min_samples_split = 10 to get smoother predictions (the tree splits a node only if it contains at least 10 samples, which is 0.1% of the 8,124 total samples). The following plot displays the most important features for the decision tree model. It can be observed that 'odor' is the most important feature, and the model is highly dependent on it. Therefore, for the decision tree, the odor feature should be the main decision rule.



The following illustrates the decision tree classification which is based on the below characteristics. Each "True" or "False" answer to these questions guides you down a specific path, eventually leading to a prediction of whether the mushroom is safe to eat or not.

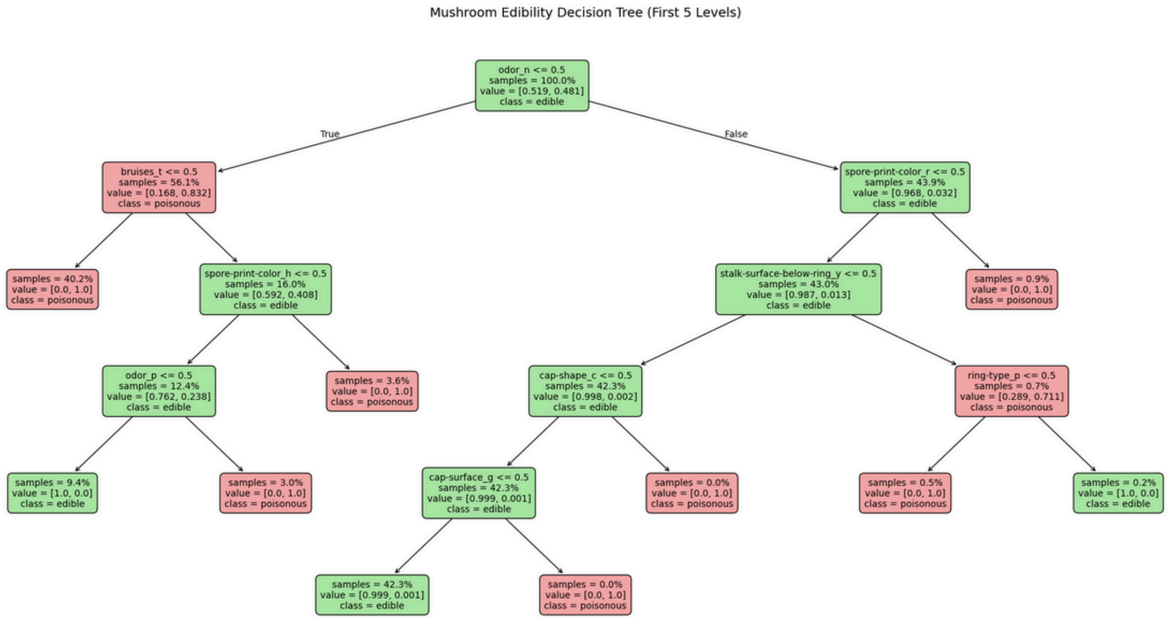
- Presence/absence of a "no odor" characteristic.
- Presence/absence of bruises.
- Specific spore print colors (chocolate, red).
- Stalk surface texture (silky).
- Cap shape (conical).
- Cap surface (gills).
- Ring type (pendant).
- Specific odor (pungent).

Here is the interpretation of the decision tree classifier.

1. Start at the Top (Root):
    - Rule: If a mushroom has no specific odor (or very low 'n' odor value) (odor_n ≤ 0.5 is True), go left.
    - Else (if it has a strong or different odor) (odor_n ≤ 0.5 is False), go right.
2. Following the Left Path (from "odor_n ≤ 0.5" is True):
    - If the mushroom has no bruises (or very few) (bruises_t ≤ 0.5 is True), it's likely POISONOUS.

   - Else (if it has bruises) (bruises_t ≤ 0.5 is False), then check its spore print color 'h' (chocolate).
     - If spore print color 'h' is low/absent (spore_print_color_h ≤ 0.5 is True), then check its odor 'p' (pungent).
       - If odor 'p' is low/absent (odor_p ≤ 0.5 is True), it's likely EDIBLE.
       - Else (if odor 'p' is present), it's likely POISONOUS.
     - Else (if spore print color 'h' is present), it's likely POISONOUS.

3. The Right Path (from "odor_n ≤ 0.5" is False):
   - Rule: If the mushroom's spore print color 'r' (red) is low/absent (spore_print_color_r ≤ 0.5 is True), then check its stalk surface below ring 'y' (silky).
     - If the stalk surface below ring 'y' is low/absent (stalk_surface-below_ring_y ≤ 0.5 is True), then check its cap shape 'c' (conical).
       - If cap shape 'c' is low/absent (cap_shape_c ≤ 0.5 is True), then check its cap surface 'g' (gills).
         - If cap surface 'g' is low/absent (cap_surface_g ≤ 0.5 is True), it's likely EDIBLE.
         - Else (if cap surface 'g' is present), it's likely POISONOUS.
       - Else (if cap shape 'c' is present), it's likely POISONOUS.
     - Else (if stalk surface below ring 'y' is present) (stalk_surface-below_ring_y ≤ 0.5 is False), then check its ring type 'p' (pendant).
       - If ring type 'p' is low/absent (ring_type_p ≤ 0.5 is True), it's likely POISONOUS.
       - Else (if ring type 'p' is present), it's likely EDIBLE.
   - Else (if spore print color 'r' is present), it's likely POISONOUS.

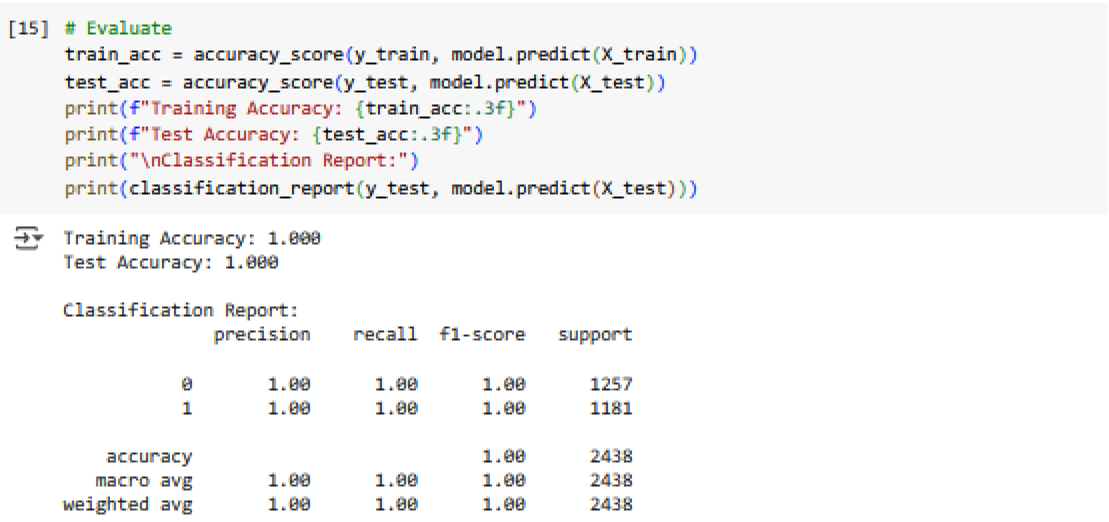


Here is the general idea as a simple first three rules,

- If a mushroom has a distinct smell (it's NOT odorless) => It's likely EDIBLE
- If a mushroom seems odorless, BUT it has bruises => It's likely POISONOUS
- If a mushroom seems odorless, AND it does NOT have bruises => It's likely EDIBLE

Accuracy of the model

It was observed that the decision tree achieved 100% accuracy on both training and test sets because the mushroom dataset contains highly predictive categorical features that clearly separate edible and poisonous classes (clean dataset). This perfect accuracy is valid in this case due to the clean, noise-free nature of the dataset and does not indicate overfitting.



```
[15] # Evaluate
     train_acc = accuracy_score(y_train, model.predict(X_train))
     test_acc = accuracy_score(y_test, model.predict(X_test))
     print(f"Training Accuracy: {train_acc:.3f}")
     print(f"Test Accuracy: {test_acc:.3f}")
     print("\nClassification Report:")
     print(classification_report(y_test, model.predict(X_test)))
```

```
Training Accuracy: 1.000
Test Accuracy: 1.000

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1257
           1       1.00      1.00      1.00      1181

    accuracy                           1.00      2438
   macro avg       1.00      1.00      1.00      2438
weighted avg       1.00      1.00      1.00      2438
```

The Decision Tree model is easy to understand for mushroom classification and gives clear insights that match real biology. It shows that mushroom safety follows strict natural rules rather than uncertain patterns. The tree's perfect results aren't a mistake; it's correctly learning the absolute biological facts in the data.

Concussion

The Decision Tree model perfectly identifies whether mushrooms are safe to eat by learning clear rules from data. It focuses on the most important clues; smells (like foul or almond scents) and colors (like green spores); just like mushroom experts do. The tree is easy to follow, with each step showing how a person would actually check a mushroom.

Since mushrooms follow strict natural patterns, the tree achieves perfect accuracy. It ignores unimportant details and focuses only on what matters. Anyone can understand how it makes decisions, and these match real field guides. For mushroom identification, Decision Trees are ideal: simple, accurate, and safe.

Source code

https://github.com/MalithaDilshan/MSC/blob/main/Computer%20Programming%20ML/Assignment/DS1501_ComputerProgrammingforDSAI_2425_DTS2412.ipynb