Edibility of Mushrooms: Methods, Findings, and Recommendations

April 1, 2025
Julia M. Cardillo
jc34142n@pace.edu
Class Name: Practical Data Science
Program Name: MS in Data Science
Seidenberg School of Computer Science and Information Systems
Pace University

Agenda

- Executive Summary
- Project Plan Recap
- Data
- Exploratory Data Analysis
- Modeling Methods
- Findings
- Business Recommendations & Technical Next Steps

Executive Summary

- **Problem:** Friendly Dog Park has a mushroom overgrowth problem. They need a method for park staff to classify any poisonous mushrooms for removal. While about only 1% of mushroom species are toxic, mushroom poisoning in dogs can cause symptoms like vomiting, kidney failure and even result in death. Treatment must be prompt, and veterinary visits can be expensive.

- **Solution:** Use a classification model to predict the edibility of mushrooms based on various physical characteristics. With this model as a guide, park staff can inspect the grounds daily prior to park opening and remove any potentially poisonous mushrooms. Guides can also be posted on billboards for park guests to reference and follow.

Project Plan Recap

Deliverable	Due Date	Status
Data & EDA	3/25/2025	Complete
Methods, Findings, and Recommendations	4/1/2025	Complete
Final Presentation	4/22/2025	In Progress

Data

Data

- Mushroom Dataset (University of California Irvine Machine Learning Repository)
- 8,124 observations (mushrooms)
- Time period: none
- Mode imputation used for one feature (stalk-root)
- Assumption: Edibility pertains to being able to be consumed by dogs.

Exploratory Data Analysis

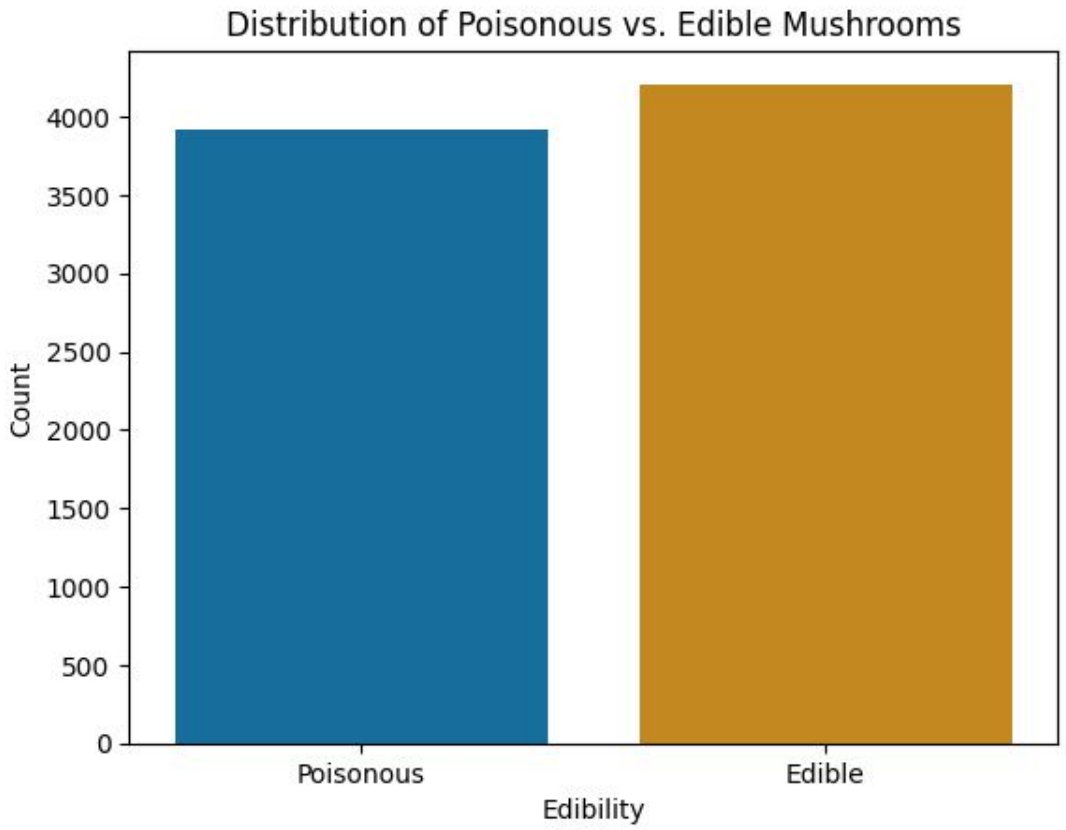
Distribution of Poisonous vs. Edible Mushrooms

Key Takeaway(s):

- Slight unbalance between the amount of edible and poisonous mushrooms in the dataset.
- Park staff should not undermine the presence of poisonous mushrooms at the park, and identify each mushroom with care and caution.

So What?

- May cause model prediction inaccuracies, by making poisonous mushrooms harder to predict.



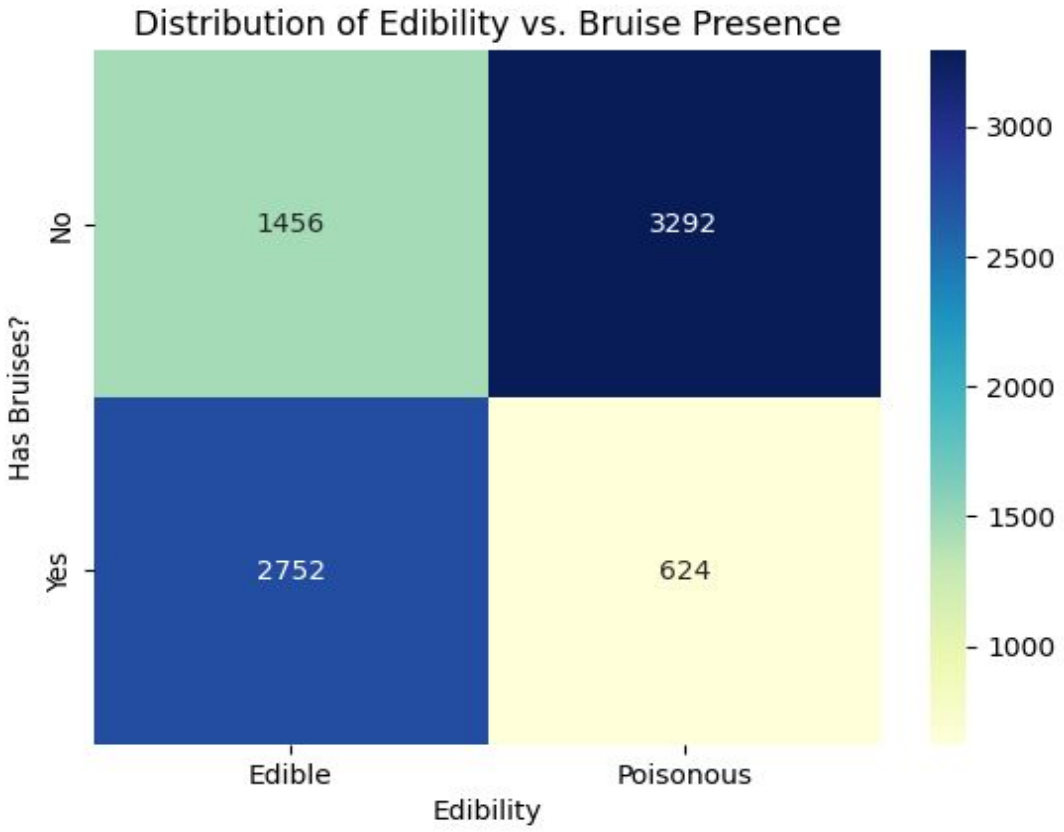
Distribution of Edibility vs. Bruise Presence

Key Takeaway(s):

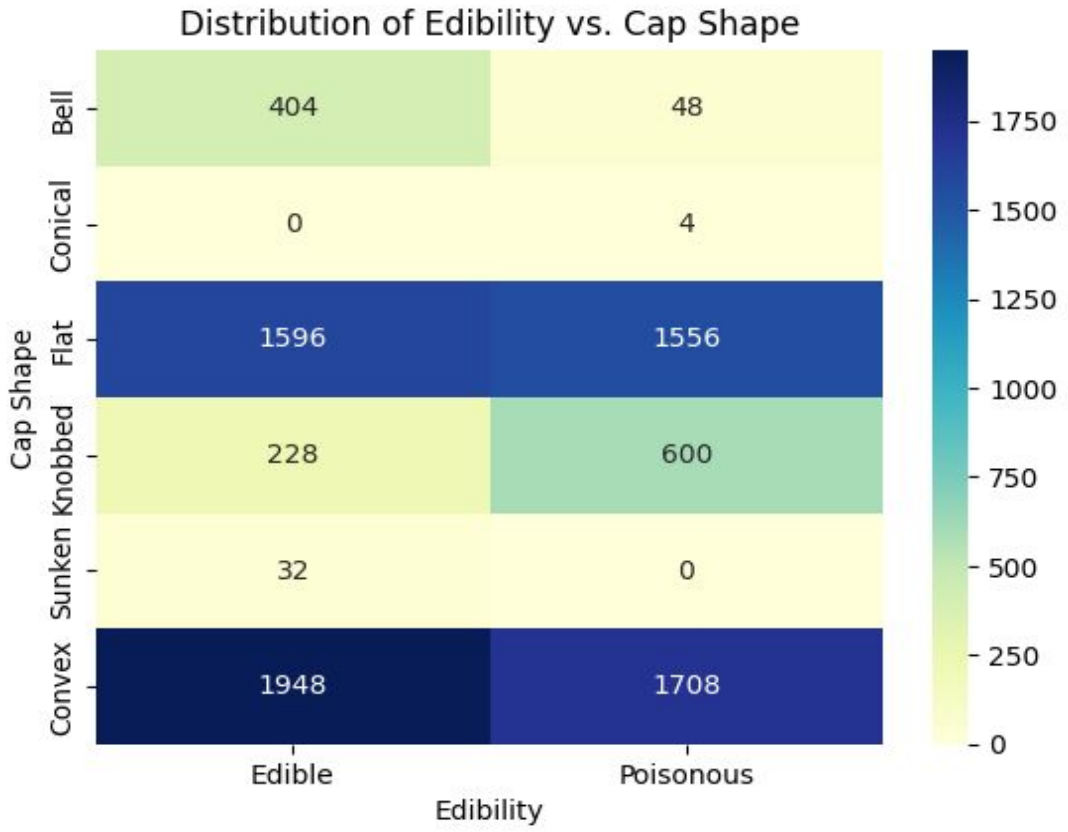
- Most of the poisonous mushrooms in the dataset did not have bruises, while most of the edible mushrooms had bruises.
- Park staff should first rule out bruising when performing inspections as it seems to be a decent indicator.

So What?

- If a mushroom does not have bruising, it is likely that it is poisonous. If a mushroom does have bruises, it is likely edible.



Distribution of Edibility vs. Cap Shape



Key Takeaway(s):

- Some cap shape types have less observations compared to others (class imbalance).
- Most mushrooms in the dataset appeared to have flat or convex cap shapes, with somewhat equal levels of poisonous and edible mushrooms for each.
- Park staff should exercise more caution when examining a mushroom's cap shape and cross reference with other indicators like bruising.

So What?

- Cap shape, at a glance, is not a glaring indicator of edibility for flat and convex cap shapes. However, convex had the highest presence of poisonous mushrooms in the dataset.

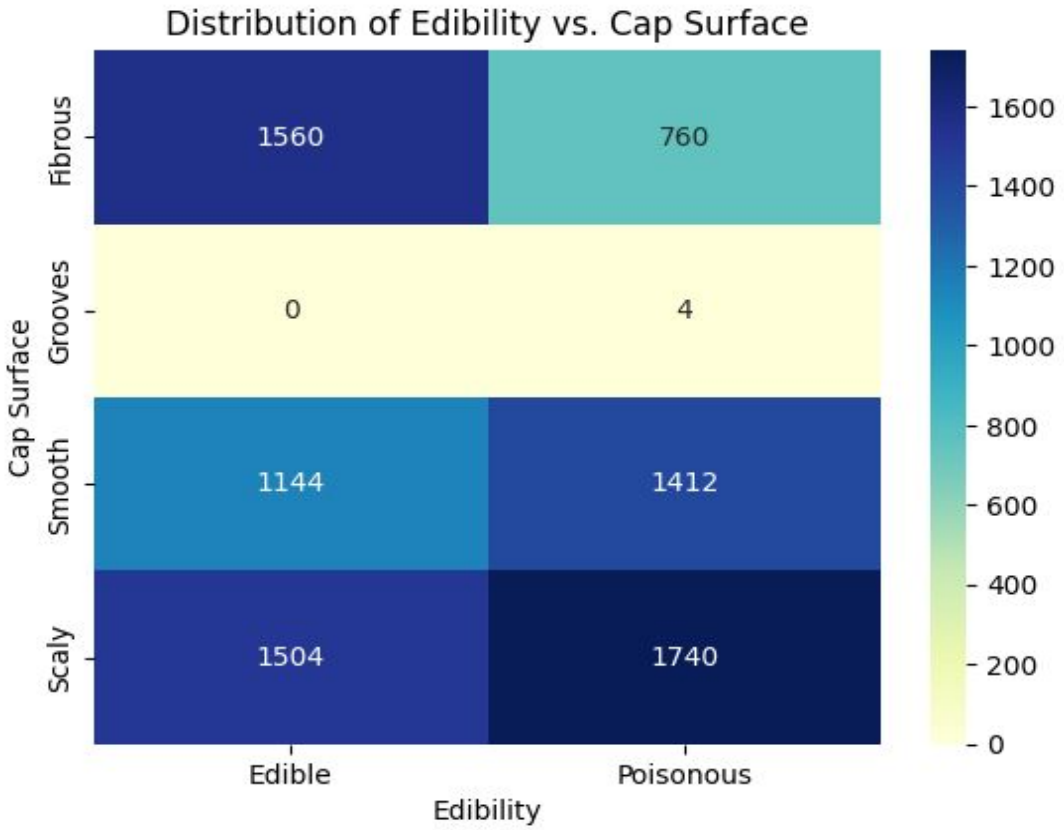
Distribution of Edibility vs. Cap Surface

Key Takeaway(s):

- Some cap surface types (grooves) lack observations compared to others.
- Scaly had the greatest number of poisonous observations, while fibrous had the least amount.
- Park staff should watch out for scaly cap surfaces in particular, due to the above point.

So What?

- Cap surface, like cap shape, is not a glaring indicator of edibility. However, a scaly cap surface type is more likely to indicate a poisonous mushroom.



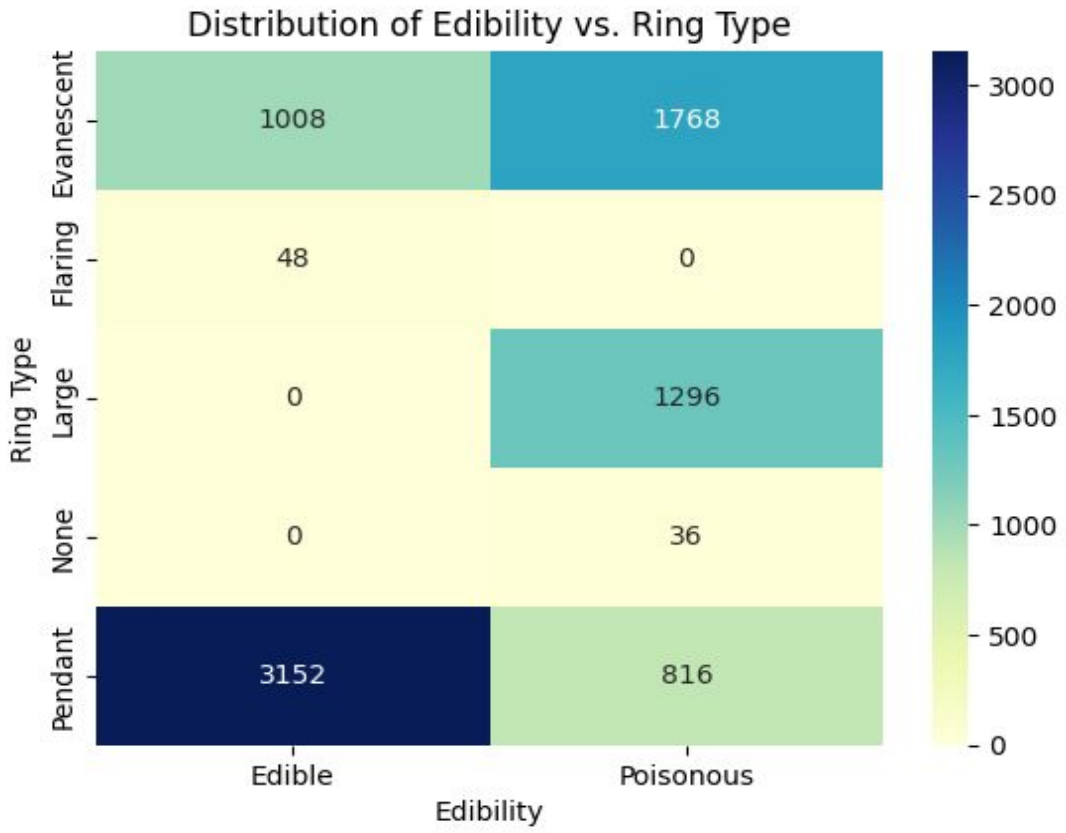
Distribution of Edibility vs. Ring Type

Key Takeaway(s):

- Most mushroom ring types in the dataset appear to be evanescent or pendant.
- Most of the edible mushrooms had a pendant ring type, while most of the poisonous mushrooms had a evanescent ring type.
- Park staff should be especially wary of evanescent ring types, and cross reference with other physical features.

So What?

- Mushrooms with a pendant ring type are more likely to be edible than other ring types, and mushrooms with an evanescent ring type are more likely to be poisonous.



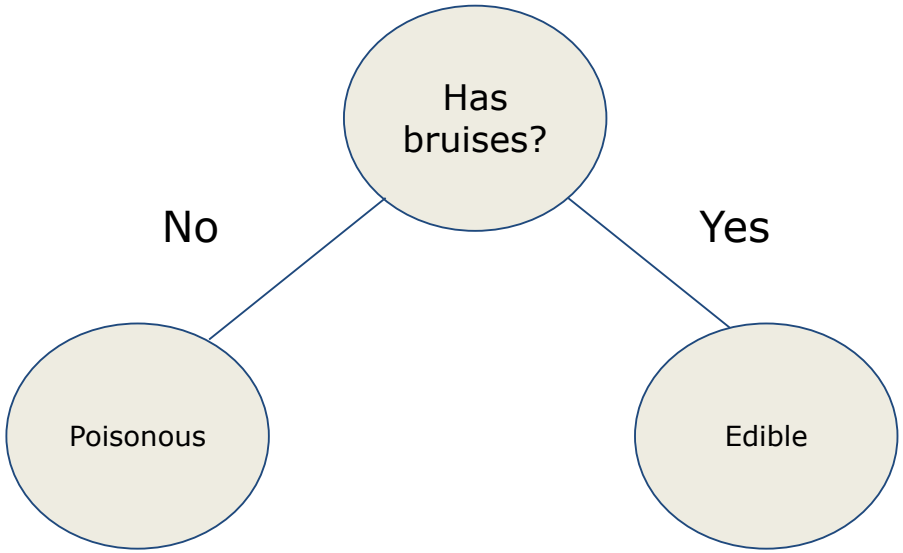
Modeling Methods

Features

- **Outcome Variable:** poisonous
  - Is the mushroom edible or poisonous?
- **Features:** 22 total features
- **After feature selection (Select K Best):** bruises, gill-spacing, gill-size, gill-color, stalk-root, stalk-surface-above-ring, ring-type, spore-print-color,population, habitat
  - Remove redundant features and improve accuracy of the model by using only the most important features.

Decision Tree Model

- Simple to understand, can handle categorical and numerical data.
- Can capture nonlinear relationships well.
- Breaks the main question down into smaller questions to answer.
- Think of "20 Questions" game, where you ask a bunch of yes or no questions in order to derive an answer.
- Technical Explanation



Findings

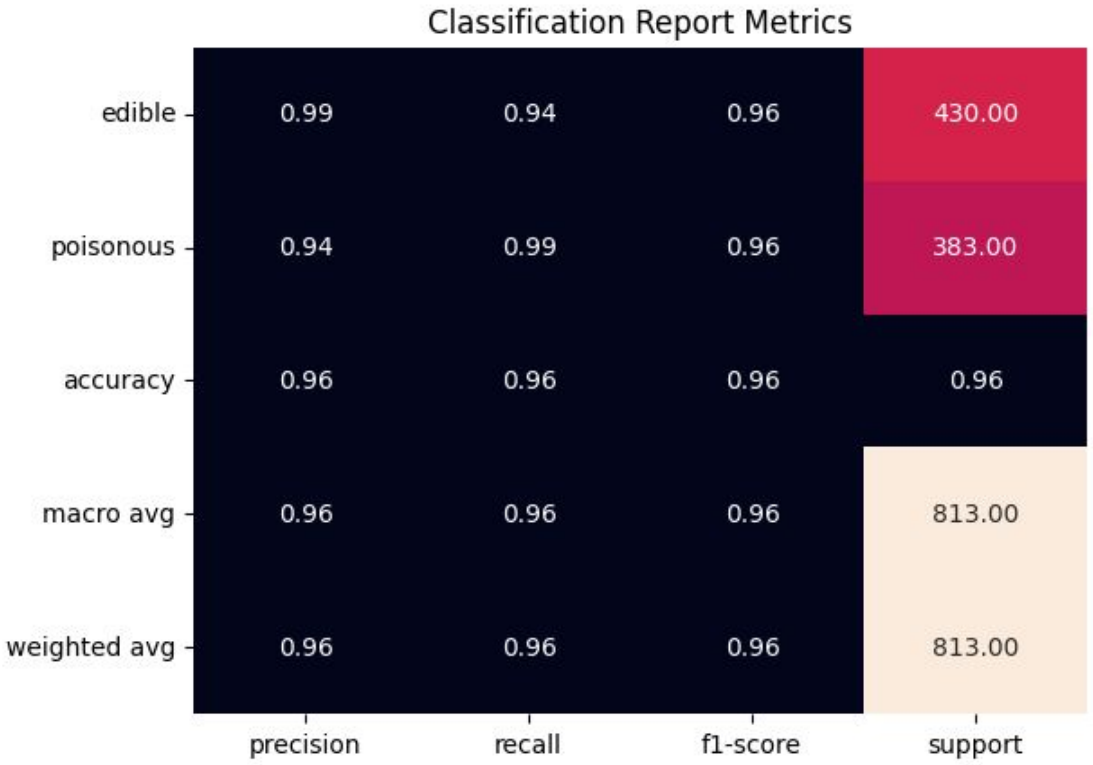
Overall Performance Evaluation

Key Takeaway(s):

- The selected physical features, in combination with a decision tree model performs well on test data.
- F1-score shows to be 96%, which shows that the model performs well overall. Provides a balanced evaluation of:
  - Out of all the existing poisonous mushrooms in the dataset, how many did we find? (Recall)
  - Out of the mushrooms we classified as poisonous, how many actually were poisonous? (Precision)

So What?

- Being that the model has a well-rounded performance, it can be used as a solid guide for park staff to classify poisonous mushrooms!



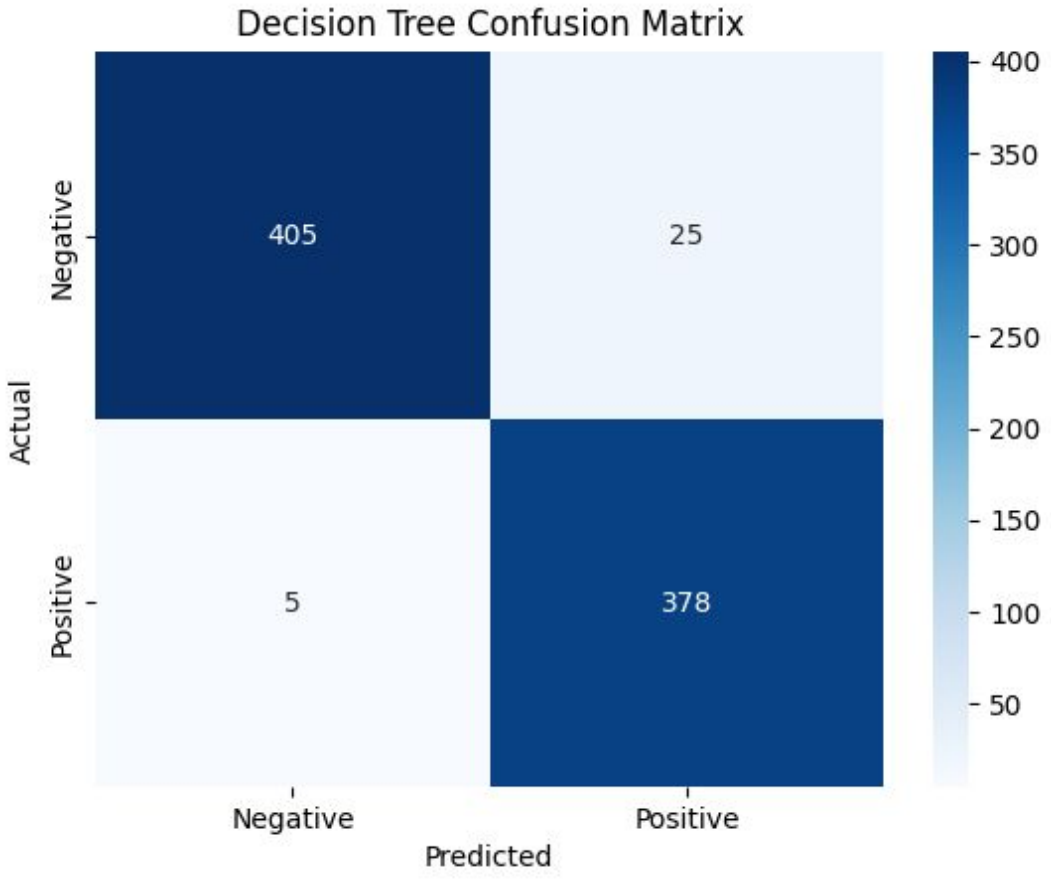
Confusion Matrix

Key Takeaway(s):

- The model does a good job of correctly classifying poisonous and edible mushrooms:
  - 25 edible mushrooms falsely predicted as poisonous.
  - 5 poisonous mushrooms falsely predicted as edible.
- "Positive" indicates poisonous.
- "Negative" indicates edible.

So What?

- Based on the confusion matrix, we can see that out of the 813 mushrooms used to test the model, only 3% of mushrooms were falsely predicted as either edible or poisonous. This, in combination with the classification report shows that the model performs well and is suitable for Friendly Dog Park to use.



Business Recommendations & Technical Next Steps

Business Recommendations

- Create a binder full of visual examples of poisonous mushroom features to train park staff.
  - Utilize the 10 features selected for the model, as these are the most important features in classifying poisonous mushrooms. They also yielded a strong model performance.
- Post visual examples on billboards around the park.
  - Similar to how guides for fishing are posted.
  - It is best to use features that show a strong trend, like mushroom bruising.
- Create an app that park patrons can download that has guides for mushroom classification, fishing, etc.

Technical Next Steps

- Use Grid Search to determine hyperparameters (parameters external to the model itself) to optimize performance.
- Create a correlation matrix between all of the features, and check for multicollinearity.

Appendix

Full Feature List

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

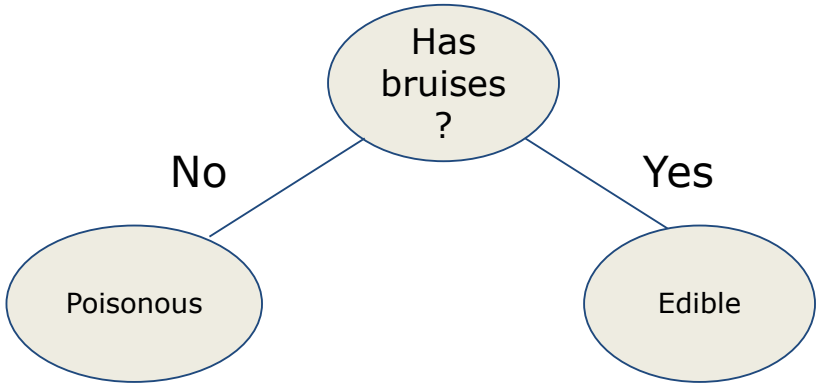
Note: to be used in the SkiKitLearn Decision Tree model, features must be encoded.

Select K Best Chi 2 Results

Feature Name: cap-shape, Chi2 Score: 105.3778, p-value: 0.0000
Feature Name: cap-surface, Chi2 Score: 214.0685, p-value: 0.0000
Feature Name: cap-color, Chi2 Score: 11.5114, p-value: 0.0007
Feature Name: bruises, Chi2 Score: 1194.2774, p-value: 0.0000
Feature Name: odor, Chi2 Score: 75.9102, p-value: 0.0000
Feature Name: gill-attachment, Chi2 Score: 3.5054, p-value: 0.0612
Feature Name: gill-spacing, Chi2 Score: 826.7953, p-value: 0.0000
Feature Name: gill-size, Chi2 Score: 1636.6068, p-value: 0.0000
Feature Name: gill-color, Chi2 Score: 5957.7645, p-value: 0.0000
Feature Name: stalk-shape, Chi2 Score: 36.5941, p-value: 0.0000
Feature Name: stalk-root, Chi2 Score: 1358.3586, p-value: 0.0000
Feature Name: stalk-surface-above-ring, Chi2 Score: 222.9824, p-value: 0.0000
Feature Name: stalk-surface-below-ring, Chi2 Score: 206.6482, p-value: 0.0000
Feature Name: stalk-color-above-ring, Chi2 Score: 119.7922, p-value: 0.0000
Feature Name: stalk-color-below-ring, Chi2 Score: 109.7894, p-value: 0.0000
Feature Name: veil-type, Chi2 Score: nan, p-value: nan
Feature Name: veil-color, Chi2 Score: 5.1268, p-value: 0.0236
Feature Name: ring-number, Chi2 Score: 25.6463, p-value: 0.0000
Feature Name: ring-type, Chi2 Score: 1950.6101, p-value: 0.0000
Feature Name: spore-print-color, Chi2 Score: 379.1327, p-value: 0.0000
Feature Name: population, Chi2 Score: 311.7667, p-value: 0.0000
Feature Name: habitat, Chi2 Score: 751.3095, p-value: 0.0000

Decision Tree Model

- Supervised classification model.
- Includes a root node, which stems into a series of decision nodes that resemble questions being asked or decisions being made. From the root nodes, leaf nodes are created that represent the results or consequences of these decisions.
- Learns patterns from data by generating if-else decision rules.
- Can handle both numerical and categorical data.
- Easy to interpret, and can be visualized.
- Can handle multi-output (multi-class target).



Notes on Validation Sets

- Validation set achieved an accuracy of 95%.
- K-fold Cross validation was also performed, with all accuracy, F1, and ROC AUC scores above 95%
- Max depth parameter was set to correct overfitting.

Sources

https://www.petmd.com/dog/poisoning/mushroom-poisoning-in-dogs

https://scikit-learn.org/stable/modules/tree.html

Github

- Click here to view the repository for this project.