classification_project

April 9, 2025

1 Mushroom Classification

Name: Gabriel Richards

Date: 9 Apr 2025

Introduction: This project analyzes the UCI Mushroom Dataset to develop classification models that can distinguish between edible and poisonous mushrooms. We'll explore various features of mushrooms and evaluate how effectively different machine learning algorithms can predict mushroom edibility. Our goal is to identify which characteristics are most predictive and which models perform best as classifiers.

1.1 1 Import & Review Data

In this section, we'll import necessary libraries and load the mushroom dataset. We'll examine the dataset structure, display the first few rows, and check for missing values. This initial inspection will help us understand the dataset characteristics and identify any data quality issues that need addressing.

To test different splits and training results, the state_setter variable allows one to set the "random_state" variable anywhere it occurs across the notebook. This allows us to test different runs of each model.

1.1.1 1.1 Load the dataset and display the first 10 rows

```
First 10 rows of the dataset:
  poisonous cap-shape cap-surface cap-color bruises odor gill-attachment  \
0         p         x           s         n       t    p               f
1         e         x           s         y       t    a               f
2         e         b           s         w       t    l               f
3         p         x           y         w       t    p               f
4         e         x           s         g       f    n               f
5         e         x           y         y       t    a               f
6         e         b           s         w       t    a               f
7         e         b           y         w       t    l               f
8         p         x           y         w       t    p               f
9         e         b           s         y       t    a               f

  gill-spacing gill-size gill-color  … stalk-surface-below-ring  \
0            c         n          k  …                        s
```

```
1            c         b          k  …                        s
2            c         b          n  …                        s
3            c         n          n  …                        s
4            w         b          k  …                        s
5            c         b          n  …                        s
6            c         b          g  …                        s
7            c         b          n  …                        s
8            c         n          p  …                        s
9            c         b          g  …                        s

  stalk-color-above-ring stalk-color-below-ring veil-type veil-color  \
0                      w                      w         p          w
1                      w                      w         p          w
2                      w                      w         p          w
3                      w                      w         p          w
4                      w                      w         p          w
5                      w                      w         p          w
6                      w                      w         p          w
7                      w                      w         p          w
8                      w                      w         p          w
9                      w                      w         p          w

  ring-number ring-type spore-print-color population habitat
0           o         p                 k          s       u
1           o         p                 n          n       g
2           o         p                 n          n       m
3           o         p                 k          s       u
4           o         e                 n          a       g
5           o         p                 k          n       g
6           o         p                 k          n       m
7           o         p                 n          s       m
8           o         p                 k          v       g
9           o         p                 k          s       m

[10 rows x 23 columns]
```

1.1.2 1.2 Check for missing values and display summary statistics

```
Missing values in each column:
poisonous                   0
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                  0
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
dtype: int64

Summary statistics for categorical data:
       poisonous cap-shape cap-surface cap-color bruises  odor  \
count       8124      8124        8124      8124    8124  8124
unique         2         6           4        10       2     9
top            e         x           y         n       f     n
freq        4208      3656        3244      2284    4748  3528

       gill-attachment gill-spacing gill-size gill-color  …  \
count             8124         8124      8124       8124  …
unique               2            2         2         12  …
top                  f            c         b          b  …
freq              7914         6812      5612       1728  …

       stalk-surface-below-ring stalk-color-above-ring stalk-color-below-ring  \
count                      8124                   8124                   8124
unique                        4                      9                      9
top                           s                      w                      w
freq                       4936                   4464                   4384

       veil-type veil-color ring-number ring-type spore-print-color  \
count       8124       8124        8124      8124              8124
unique         1          4           3         5                 9
top            p          w           o         p                 w
freq        8124       7924        7488      3968              2388

       population habitat
count        8124    8124
unique          6       7
top             v       d
freq         4040    3148
```

```
[4 rows x 23 columns]

Unique values in each column:
poisonous: ['p' 'e']
cap-shape: ['x' 'b' 's' 'f' 'k' 'c']
cap-surface: ['s' 'y' 'f' 'g']
cap-color: ['n' 'y' 'w' 'g' 'e' 'p' 'b' 'u' 'c' 'r']
bruises: ['t' 'f']
odor: ['p' 'a' 'l' 'n' 'f' 'c' 'y' 's' 'm']
gill-attachment: ['f' 'a']
gill-spacing: ['c' 'w']
gill-size: ['n' 'b']
gill-color: ['k' 'n' 'g' 'p' 'w' 'h' 'u' 'e' 'b' 'r' 'y' 'o']
stalk-shape: ['e' 't']
stalk-root: ['e' 'c' 'b' 'r' '?']
stalk-surface-above-ring: ['s' 'f' 'k' 'y']
stalk-surface-below-ring: ['s' 'f' 'y' 'k']
stalk-color-above-ring: ['w' 'g' 'p' 'n' 'b' 'e' 'o' 'c' 'y']
stalk-color-below-ring: ['w' 'p' 'g' 'b' 'n' 'e' 'y' 'o' 'c']
veil-type: ['p']
veil-color: ['w' 'n' 'o' 'y']
ring-number: ['o' 't' 'n']
ring-type: ['p' 'e' 'l' 'f' 'n']
spore-print-color: ['k' 'n' 'u' 'h' 'w' 'r' 'o' 'y' 'b']
population: ['s' 'n' 'a' 'v' 'y' 'c']
habitat: ['u' 'g' 'm' 'd' 'p' 'w' 'l']
```

Reflection 1: What do you notice about the dataset? Are there any data issues?

- The dataset contains 8,124 mushroom samples with 23 features, all of which are categorical
- There's a relatively balanced distribution between poisonous/edible (4,208 edible vs 3,916 poisonous mushrooms)
- All of the features are categorical. They are all encoded as single letters, and will require transformation before modeling
- There are missing values in the 'stalk-root' feature, encoded as '?' (2,480 total)
- After dropping rows with missing values, the dataset size reduced to 5,644 samples

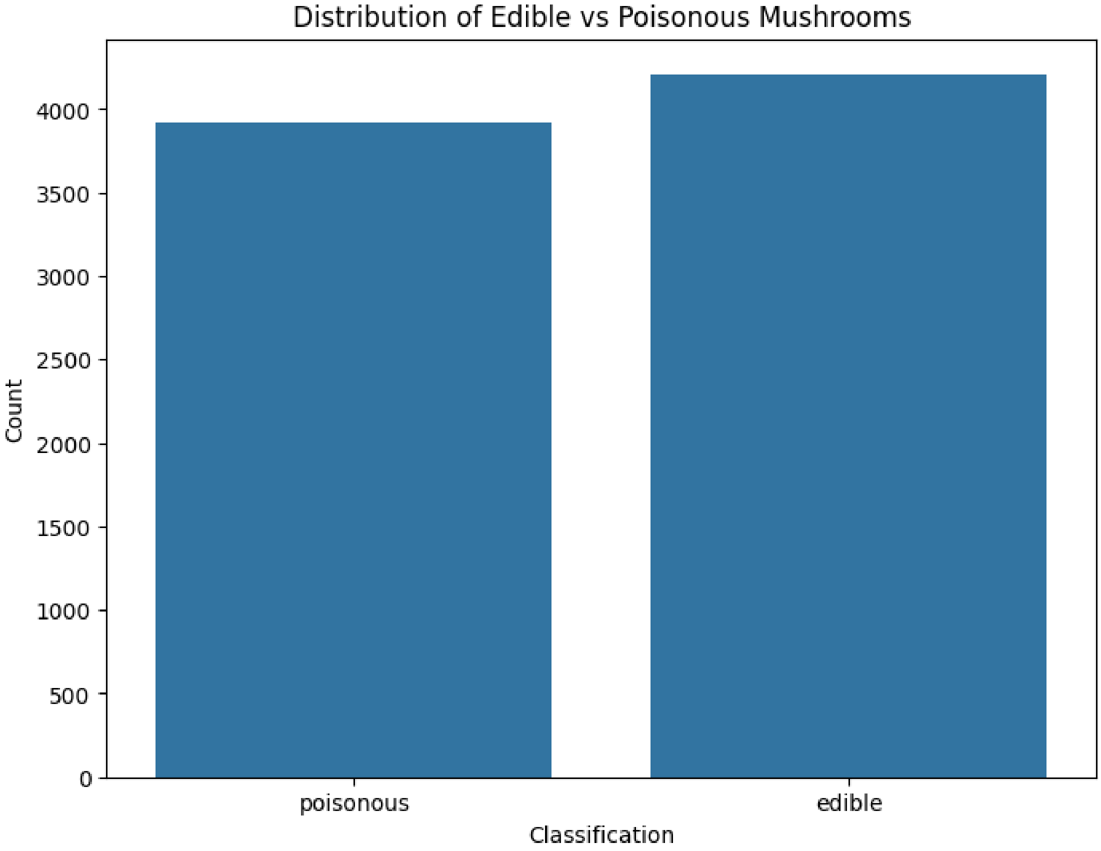
1.2 2 Data Exploration and Preparation

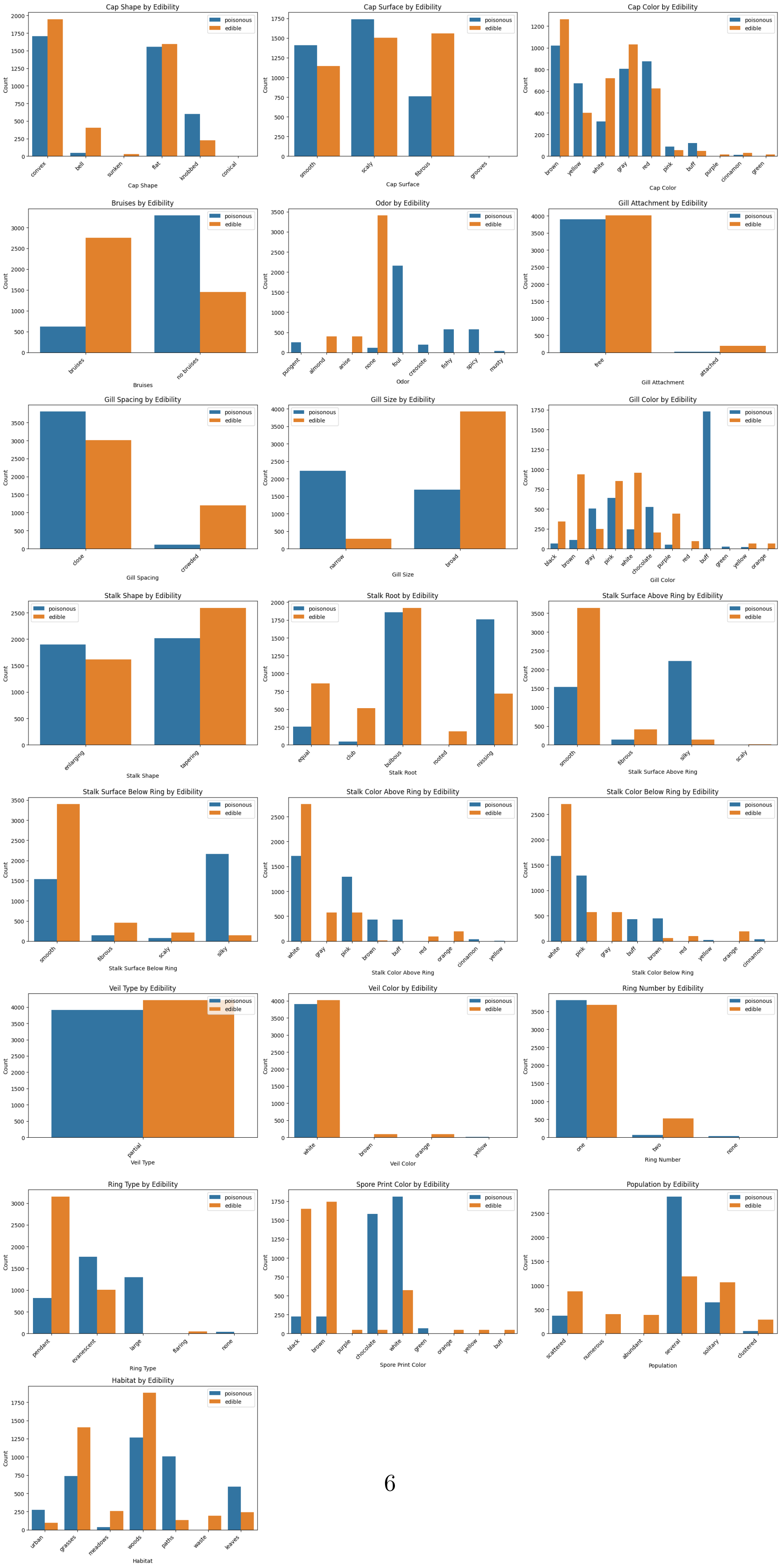
Here we'll visualize and analyze the distributions of various mushroom features, examining how they relate to edibility. We'll create plots to identify patterns and relationships, particularly looking for features that show clear separation between edible and poisonous classes. This exploration will guide our feature selection for classification models.

1.2.1 2.1 Explore data patterns and distributions

First, let's set up a mapping to clarify which single letters in the data set map to which adjectives. This is helpful for reference and for any visual depictions

Before we pick our features, it would be handy to visualize some basic details about the set. We are going to set whether the mushroom is edible or poisonous as our target variable, so let's look at the total distribution of that across the set. Let's also look at how each of the categories breaks down into subcategories in terms of edible and poisonous.




Cap Shape by Edibility
poisonous
edible
Count
0
250
500
750
1000
1250
1500
1750
2000
convex
bell
sunken
flat
knobbed
conical
Cap Shape
Cap Surface by Edibility
poisonous
edible
Count
smooth
scaly
fibrous
grooves
Cap Surface
Cap Color by Edibility
poisonous
edible
Count
brown
yellow
white
gray
red
pink
buff
purple
cinnamon
green
Cap Color
Bruises by Edibility
poisonous
edible
Count
bruises
no bruises
Bruises
Odor by Edibility
poisonous
edible
Count
pungent
almond
anise
none
foul
creosote
fishy
spicy
musty
Odor
Gill Attachment by Edibility
poisonous
edible
Count
free
attached
Gill Attachment
Gill Spacing by Edibility
poisonous
edible
Count
close
crowded
Gill Spacing
Gill Size by Edibility
poisonous
edible
Count
narrow
broad
Gill Size
Gill Color by Edibility
poisonous
edible
Count
black
brown
gray
pink
white
chocolate
purple
red
buff
green
yellow
orange
Gill Color
Stalk Shape by Edibility
poisonous
edible
Count
enlarging
tapering
Stalk Shape
Stalk Root by Edibility
poisonous
edible
Count
equal
club
bulbous
rooted
missing
Stalk Root
Stalk Surface Above Ring by Edibility
poisonous
edible
Count
smooth
fibrous
silky
scaly
Stalk Surface Above Ring
Stalk Surface Below Ring by Edibility
poisonous
edible
Count
smooth
fibrous
scaly
silky
Stalk Surface Below Ring
Stalk Color Above Ring by Edibility
poisonous
edible
Count
white
gray
pink
brown
buff
red
orange
cinnamon
yellow
Stalk Color Above Ring
Stalk Color Below Ring by Edibility
poisonous
edible
Count
white
pink
gray
buff
brown
red
yellow
orange
cinnamon
Stalk Color Below Ring
Veil Type by Edibility
poisonous
edible
Count
partial
Veil Type
Veil Color by Edibility
poisonous
edible
Count
white
brown
orange
yellow
Veil Color
Ring Number by Edibility
poisonous
edible
Count
one
two
none
Ring Number
Ring Type by Edibility
poisonous
edible
Count
pendant
evanescent
large
flaring
none
Ring Type
Spore Print Color by Edibility
poisonous
edible
Count
black
brown
purple
chocolate
white
green
orange
yellow
buff
Spore Print Color
Population by Edibility
poisonous
edible
Count
scattered
numerous
abundant
several
solitary
clustered
Population
Habitat by Edibility
poisonous
edible
Count
urban
grasses
meadows
woods
paths
waste
leaves
Habitat

Lastly, before we proceed, let's clean the data for any null or missing values.

```
Checking for '?' values in the data:
stalk-root has missing values encoded as '?'
Dataset shape after dropping missing values: (5644, 23)
Data after encoding categorical variables:
   poisonous  cap-shape  cap-surface  cap-color  bruises  odor  \
0          1          5            2          4        1     6
1          0          5            2          7        1     0
2          0          0            2          6        1     3
3          1          5            3          6        1     6
4          0          5            2          3        0     5

   gill-attachment  gill-spacing  gill-size  gill-color  …  \
0                1             0          1           2  …
1                1             0          0           2  …
2                1             0          0           3  …
3                1             0          1           3  …
4                1             1          0           2  …

   stalk-surface-below-ring  stalk-color-above-ring  stalk-color-below-ring  \
0                         2                       5                       5
1                         2                       5                       5
2                         2                       5                       5
3                         2                       5                       5
4                         2                       5                       5

   veil-type  veil-color  ring-number  ring-type  spore-print-color  \
0          0           0            1          3                  1
1          0           0            1          3                  2
2          0           0            1          3                  2
3          0           0            1          3                  1
4          0           0            1          0                  2

   population  habitat
0           3        5
1           2        1
2           2        3
3           3        5
4           0        1

[5 rows x 23 columns]
```

Reflection 2: What patterns or anomalies do you see? Do any features stand out? What preprocessing steps were necessary to clean and improve the data?

- Certain features show clear separation between edible and poisonous mushrooms:
  - Odor is highly predictive - mushrooms with “foul,” “fishy,” “spicy,” and “pungent” odors are almost exclusively poisonous
  - Spore print color shows strong patterns - “chocolate” spore prints are consistently poisonous
- To preprocess them, I:
  - Converted ‘?’ values to NaN and dropped those rows
  - Used LabelEncoder to transform categorical variables to numeric values
  - Created mapping dictionaries for all features to maintain interpretability

1.3 3 Feature Selection & Definition

Using our examination above, we can select features to train our classifiers on.

1.3.1 3.1 Choose features and target

Based on our exploratory analysis, I’ve selected four different feature sets to test. Two of our feature sets, or cases, are individual features and the other two are pairs: - odor alone, - spore print color alone, - spore print color with gill color, - and bruises with habitat.

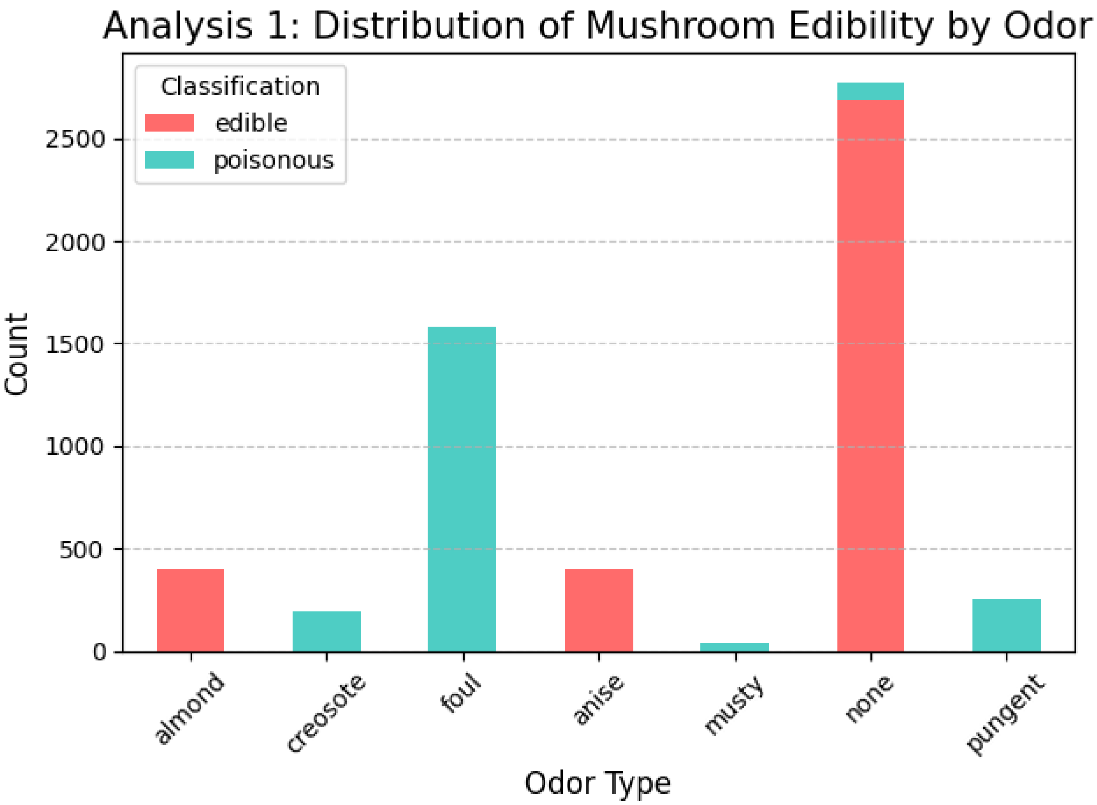
- Each set was chosen to test different biological aspects of mushrooms and their predictive power. At the end of the day, we want to know whether we can train a model to accurately predict whether a mushroom is poisonous when fed our features. This means our target variable is ‘poisonous’, a binary classification of edible (0) or poisonous (1) when consumed by humans.

1.3.2 3.2 Define X and y

The code below maps each of our cases/analyses to variables we will used to split and train models on the data.

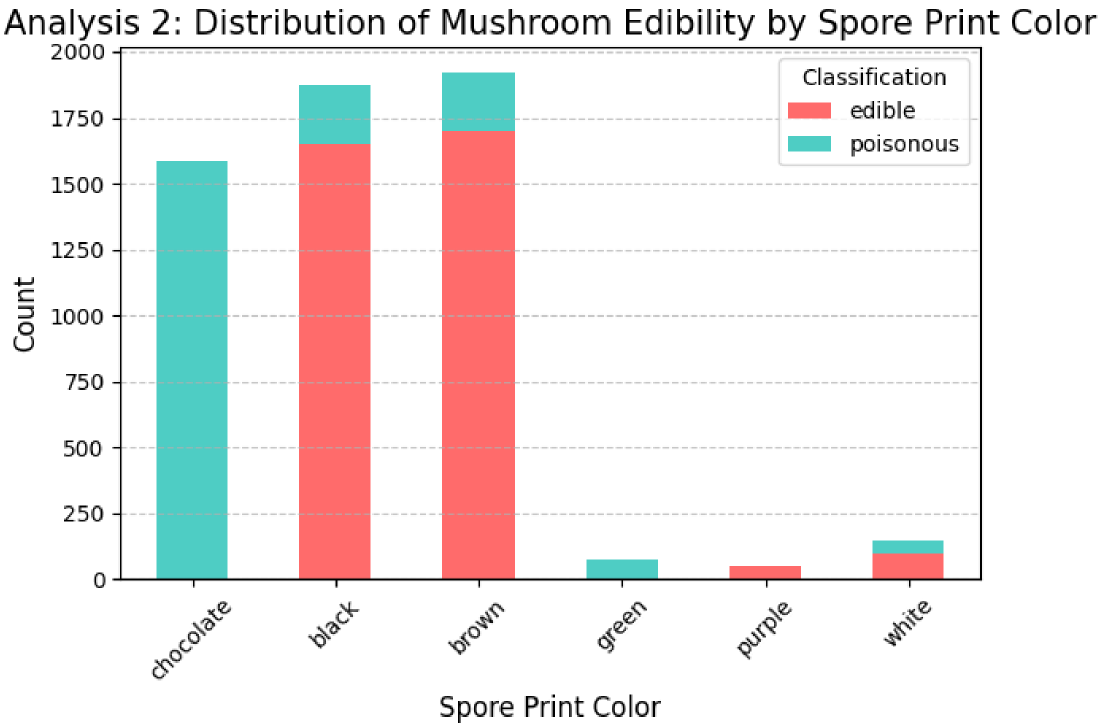
```
Analysis 1 - Odor only:
Features shape: (5644, 1)
   odor
0     6
1     0
2     3
3     6
4     5

===== ANALYSIS 1: ODOR =====
'odor' has 7 different categories:
- n (none): 2776 instances
- f (foul): 1584 instances
- a (almond): 400 instances
- l (anise): 400 instances
- p (pungent): 256 instances
- c (creosote): 192 instances
- m (musty): 36 instances
```

```
<Figure size 1200x600 with 0 Axes>
```



```
Analysis 2 - Spore print color only:
Features shape: (5644, 1)
   spore-print-color
0                  1
1                  2
2                  2
3                  1
4                  2

===== ANALYSIS 2: SPORE PRINT COLOR =====
'spore-print-color' has 6 different categories:
- n (brown): 1920 instances
- k (black): 1872 instances
- h (chocolate): 1584 instances
- w (white): 148 instances
- r (green): 72 instances
- u (purple): 48 instances
```

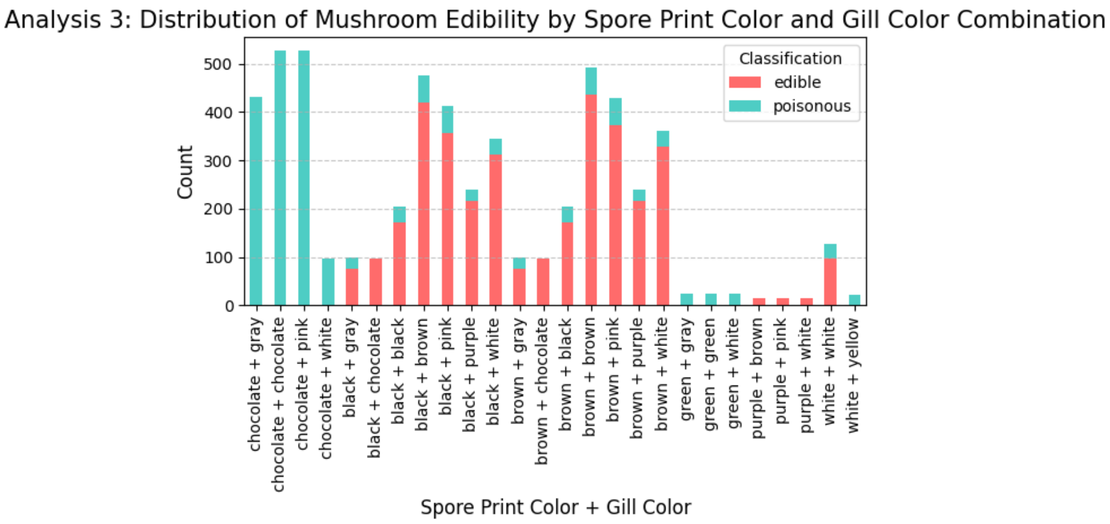
```
<Figure size 1200x600 with 0 Axes>
```



```
Analysis 3 - Spore Print Color + Gill Color:
Features shape: (5644, 2)
   spore-print-color  gill-color
0                  1           2
1                  2           2
2                  2           3
3                  1           3
4                  2           2

===== ANALYSIS 3: SPORE PRINT COLOR AND GILL COLOR =====
'spore-print-color' has 6 subcategories:
- n (brown): 1920 instances
- k (black): 1872 instances
- h (chocolate): 1584 instances
- w (white): 148 instances
- r (green): 72 instances
- u (purple): 48 instances

'gill-color' has 9 subcategories:
```

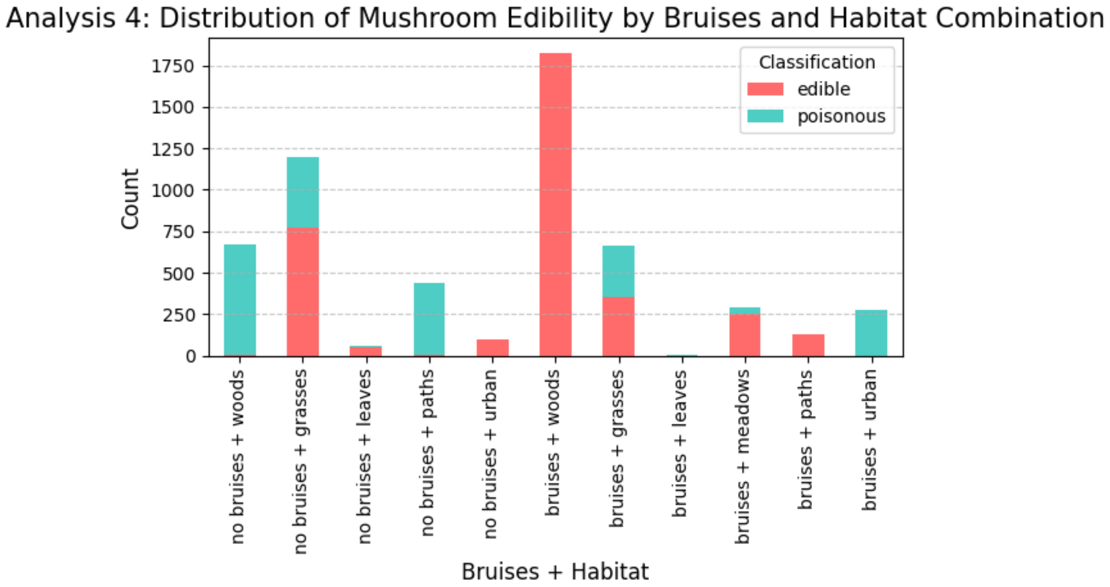
```
- p (pink): 1384 instances
- n (brown): 984 instances
- w (white): 966 instances
- h (chocolate): 720 instances
- g (gray): 656 instances
- u (purple): 480 instances
- k (black): 408 instances
- r (green): 24 instances
- y (yellow): 22 instances

Total possible combinations: 54
Actual combinations found in data: 26

Actual combinations:
- h (chocolate) + g (gray): 432 instances
- h (chocolate) + h (chocolate): 528 instances
- h (chocolate) + p (pink): 528 instances
- h (chocolate) + w (white): 96 instances
- k (black) + g (gray): 100 instances
- k (black) + h (chocolate): 96 instances
- k (black) + k (black): 204 instances
- k (black) + n (brown): 476 instances
- k (black) + p (pink): 412 instances
- k (black) + u (purple): 240 instances
- k (black) + w (white): 344 instances
- n (brown) + g (gray): 100 instances
- n (brown) + h (chocolate): 96 instances
- n (brown) + k (black): 204 instances
- n (brown) + n (brown): 492 instances
- n (brown) + p (pink): 428 instances
- n (brown) + u (purple): 240 instances
- n (brown) + w (white): 360 instances
- r (green) + g (gray): 24 instances
- r (green) + r (green): 24 instances
- r (green) + w (white): 24 instances
- u (purple) + n (brown): 16 instances
- u (purple) + p (pink): 16 instances
- u (purple) + w (white): 16 instances
- w (white) + w (white): 126 instances
- w (white) + y (yellow): 22 instances
```

```
<Figure size 1400x800 with 0 Axes>
```



```
Analysis 4 - Bruises + Habitat:
Features shape: (5644, 2)
   bruises  habitat
0        1        5
1        1        1
2        1        3
3        1        5
4        0        1

===== ANALYSIS 4: BRUISES AND HABITAT =====
'bruises' has 2 subcategories:
- t (bruises): 3184 instances
- f (no bruises): 2460 instances

'habitat' has 6 subcategories:
- d (woods): 2492 instances
- g (grasses): 1860 instances
- p (paths): 568 instances
- u (urban): 368 instances
- m (meadows): 292 instances
- l (leaves): 64 instances

Total possible combinations: 12
Actual combinations found in data: 11

Actual combinations:
- f (no bruises) + d (woods): 668 instances
- f (no bruises) + g (grasses): 1200 instances
- f (no bruises) + l (leaves): 56 instances
- f (no bruises) + p (paths): 440 instances
- f (no bruises) + u (urban): 96 instances
- t (bruises) + d (woods): 1824 instances
- t (bruises) + g (grasses): 660 instances
- t (bruises) + l (leaves): 8 instances
- t (bruises) + m (meadows): 292 instances
- t (bruises) + p (paths): 128 instances
- t (bruises) + u (urban): 272 instances
```

```
<Figure size 1400x800 with 0 Axes>
```



Reflection 3: Why did you choose these features? How might they impact predictions or accuracy? Case by case: * Odor (Case 1): Visual inspection showed clear separation between classes. Smell is often sign something is wrong in nature! * Spore print color (Case 2): Shows strong correlation with edibility; could also have biological signifance. The hope is that color patterns like “chocolate” being poisonous and “brown” being edible enable the model to make clean classifications. * Spore print color + gill color (Case 3): Spores are typically emitted from the gills on the underside of a mushroom. This will help us see if there is a trend between gills being a color, which produce specific spore print colors, which are aligned with being poisonous or edible * Bruises + habitat (Case 4): This combination might reveal whether certain environments promote toxicity in mushrooms that bruise

1.4 4 Data Splitting and First Model Training

Now that we have mapped our target variable and features, we can go on to split the data set to train and test each model.

1.4.1 4.1 Split the data into training and test sets

For each of our four feature cases, we'll split the data into training (80%) and testing (20%) sets. For our first run through, we'll use these classifier models for each case: 1. Decision Tree for odor 2. Decision Tree for spore print color 3. Random Forest for spore print color + gill color 4. Random Forest for bruises + habitat

```
Case 1 Training set size: 4515
Case 1 Test set size: 1129

Case 2 Training set size: 4515
Case 2 Test set size: 1129

Case 3 Training set size: 4515
Case 3 Test set size: 1129

Case 4 Training set size: 4515
Case 4 Test set size: 1129
```

This confirms the splitter is working as expected, with 80% going into the Training set, and 20% going into the test set.

1.4.2 4.2 Train Classifiers

With the data split, we can proceed to train the model on each case using the 80% of each split set aside for training.

4.2.1 Odor - Decision Tree

```
DecisionTreeClassifier(random_state=529)
```

4.2.2 Spore Print Color - Decision Tree

```
DecisionTreeClassifier(random_state=529)
```

4.2.3 Spore Print Color + Gill Color - Random Forest

```
RandomForestClassifier(random_state=529)
```

4.2.4 Bruises + Habitat - Random Forest

```
RandomForestClassifier(random_state=529)
```

1.4.3 4.3 Evaluate performance

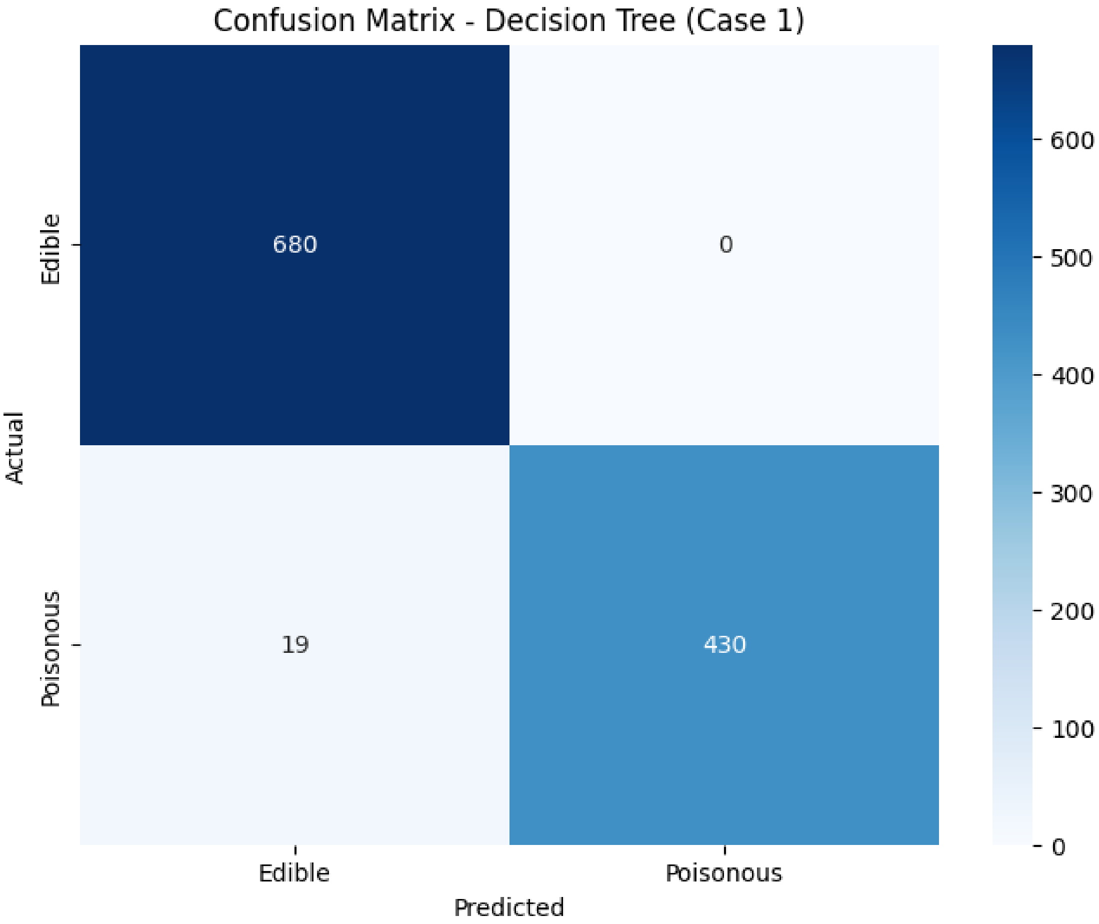
Now let's take a look at the metrics to determine how each one performed. We'll assess model performance using accuracy, precision, recall, and F1 score metrics with a summary at the end.

Confusion matrices help us visualize prediction results, showing true positives, false positives, true negatives, and false negatives.

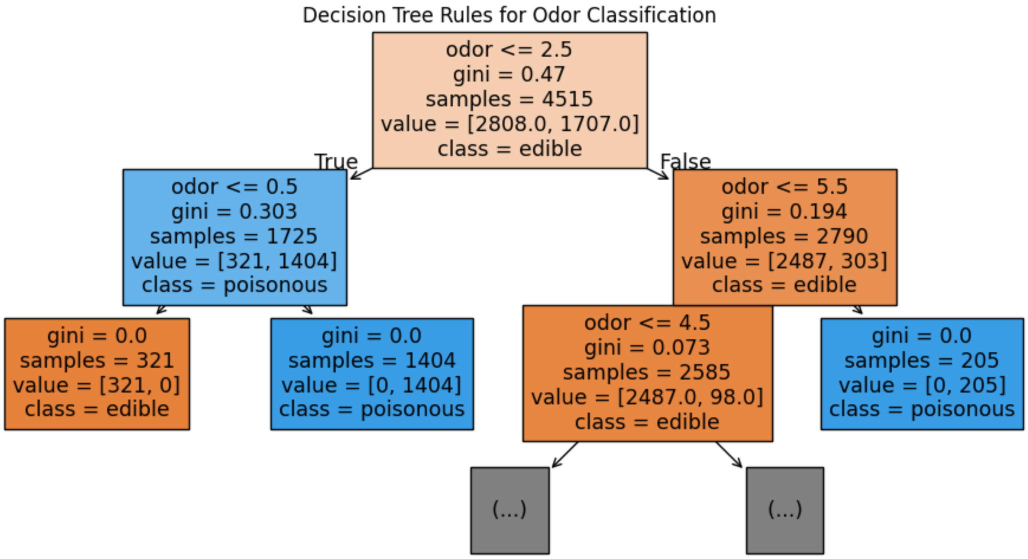
It's wise to pay special attention to false negatives (poisonous mushrooms misclassified as edible) since these represent the most dangerous errors.

4.3.1 Odor - Decision Tree Performance

```
Decision Tree Model Performance (Case 1 - Odor only):
Accuracy: 0.9832
Precision: 1.0000
Recall: 0.9577
F1 Score: 0.9784
```



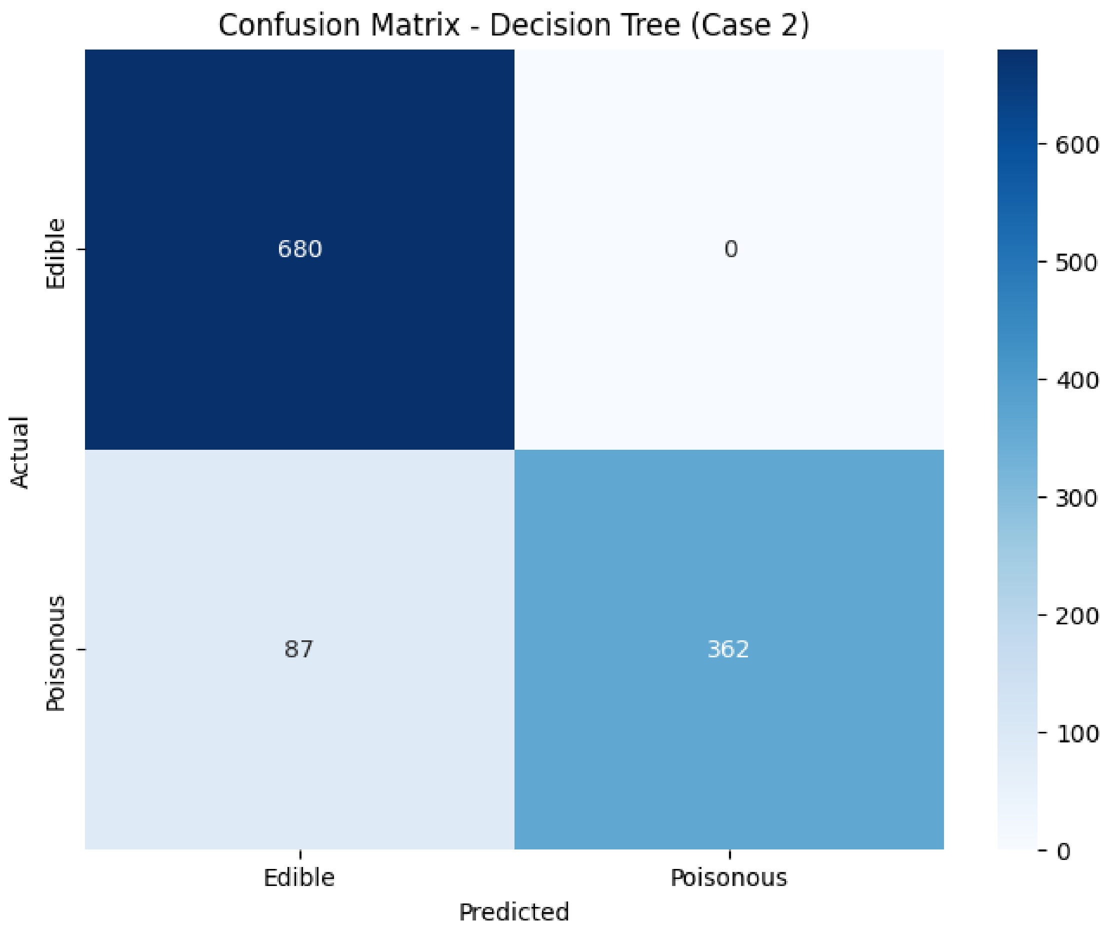
```
==== Validation for Case 1 (Odor) Decision Tree ====
Test set class distribution: [680 449]
```



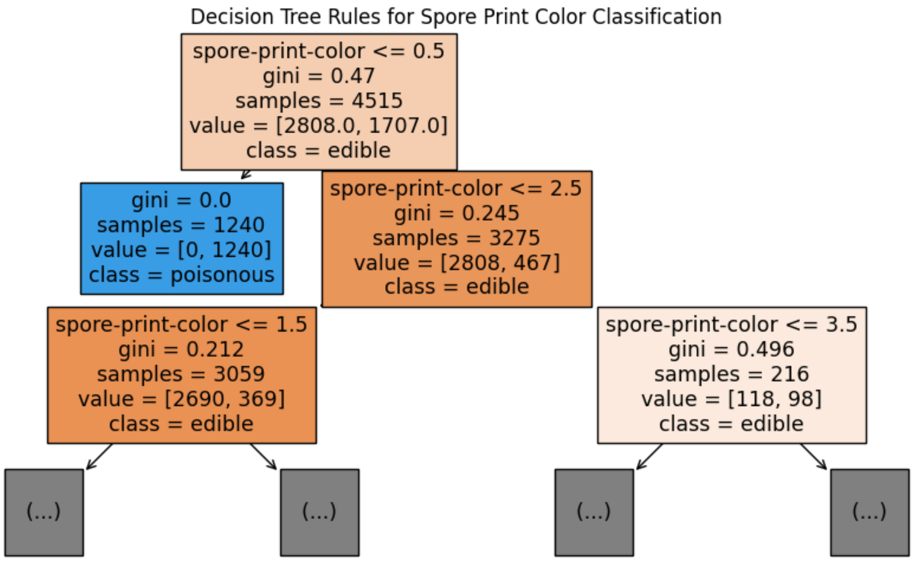
```
Sample prediction results (first 10):
      Actual  Predicted
4111       1          1
3324       0          0
1802       0          0
4427       1          1
1260       1          1
3106       0          0
871        0          0
2775       0          0
1397       0          0
1135       0          0
Number of misclassified samples: 19 out of 1129
```

4.3.2 Spore Print Color - Decision Tree Performance

```
Decision Tree Model Performance (Case 2 - Spore Print Color only):
Accuracy: 0.9229
Precision: 1.0000
Recall: 0.8062
F1 Score: 0.8927
```



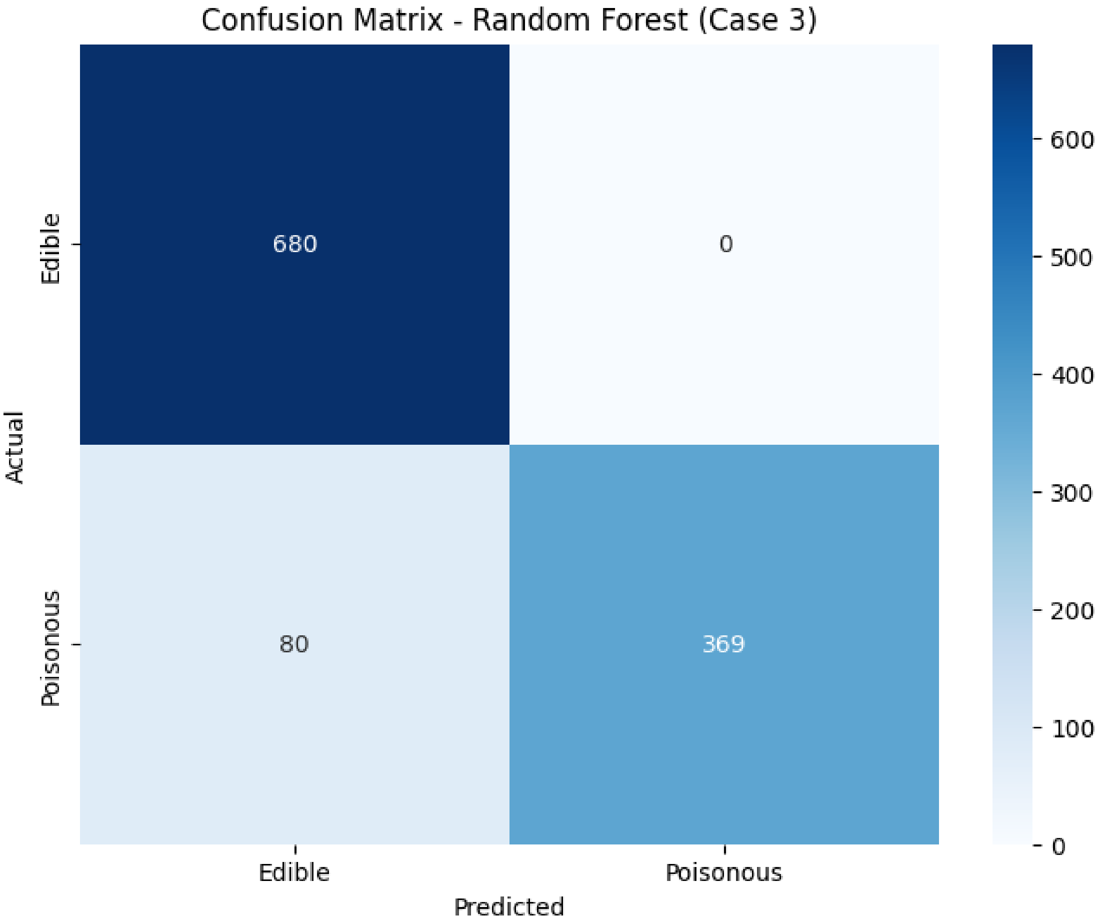
```
==== Validation for Case 2 (Spore Print Color) Decision Tree ====
Test set class distribution: [680 449]
```



```
Sample prediction results (first 10):
      Actual  Predicted
4111       1          1
3324       0          0
1802       0          0
4427       1          1
1260       1          0
3106       0          0
871        0          0
2775       0          0
1397       0          0
1135       0          0
Number of misclassified samples: 87 out of 1129
```

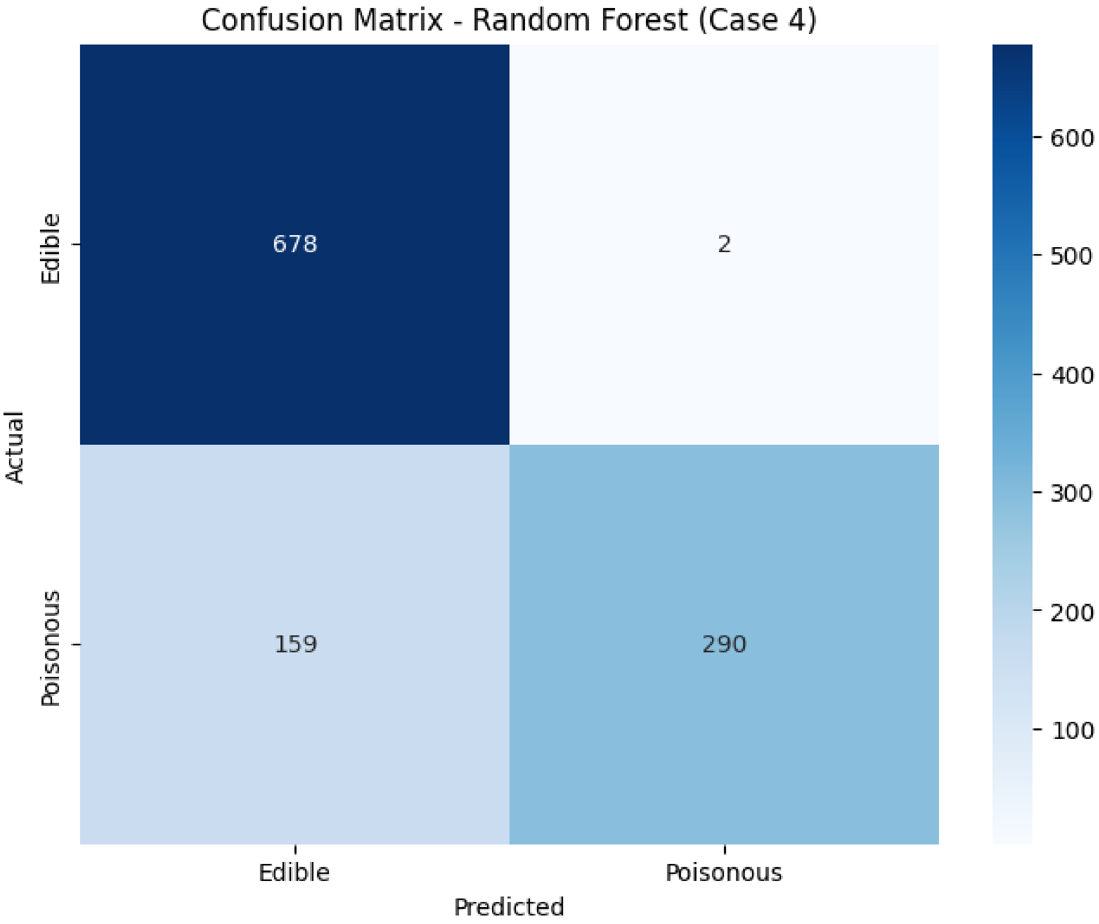
4.3.3 Spore Print Color + Gill Color - Random Forest Performance

```
Random Forest Model Performance (Case 3 - Spore Print Color + Gill Color):
Accuracy: 0.9291
Precision: 1.0000
Recall: 0.8218
F1 Score: 0.9022
```



4.3.4 Bruises + Habitat - Random Forest Performance

```
Random Forest Model Performance (Case 4 - Bruises + Habitat):
Accuracy: 0.8574
Precision: 0.9932
Recall: 0.6459
F1 Score: 0.7827
```



Reflection 4: How well did the model perform? Any surprises in the results?

- All models performed remarkably well, with accuracies ranging from 85.74-98.32%
- Using Decision Tree with only odor achieved the highest accuracy (98.32%)
- Much to my surprise, there were zero false negatives in Cases 1 and 2 - the models never misclassified a poisonous mushroom as edible
  - This occurred across random seeds entered
- Even simple single-feature models achieved high accuracy, suggesting some features are extremely predictive
- The models exhibited perfect precision (1.0) in several cases. When they predicted “poisonous,” they were always correct

1.5 5 Model Comparison

1.5.1 5.1 Train Alternate Models

To compare classifiers, let’s train a second model for each feature set: 1. Logistic Regression for odor 2. Random Forest for spore print color 3. Logistic Regression for spore print color + gill color 4. Decision Tree for bruises + habitat

This will help us determine whether certain algorithms are better suited for specific feature combinations. We will examine the results at the end of the section rather than throughout it.

5.1.1 Odor - Logistic Regression Training

5.1.2 Spore Print Color - Random Forest Training

```
Number of different predictions between Decision Tree and Random Forest: 0 out
of 1129
WARNING: Models producing identical predictions. Trying different Random Forest
configuration…
Number of different predictions after adjustment: 0 out of 1129
```

5.1.3 Spore Print Color + Gill Color - Logistic Regression Training

5.1.4 Bruises + Habitat - Decision Tree Training

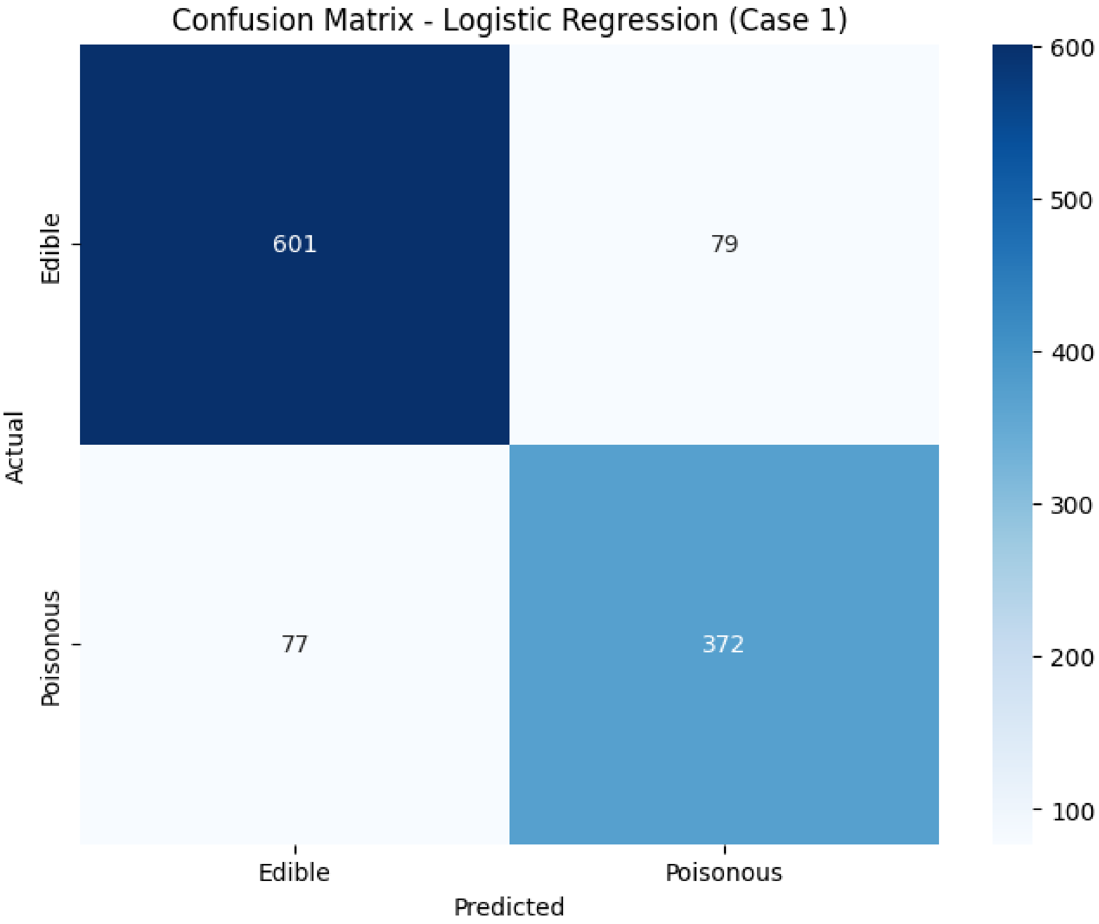
```
Number of different predictions between Random Forest and Decision Tree: 0 out
of 1129
WARNING: Models producing identical predictions. Trying different Decision Tree
configuration…
Number of different predictions after adjustment: 0 out of 1129
```

1.5.2 5.2 Compare performance of models

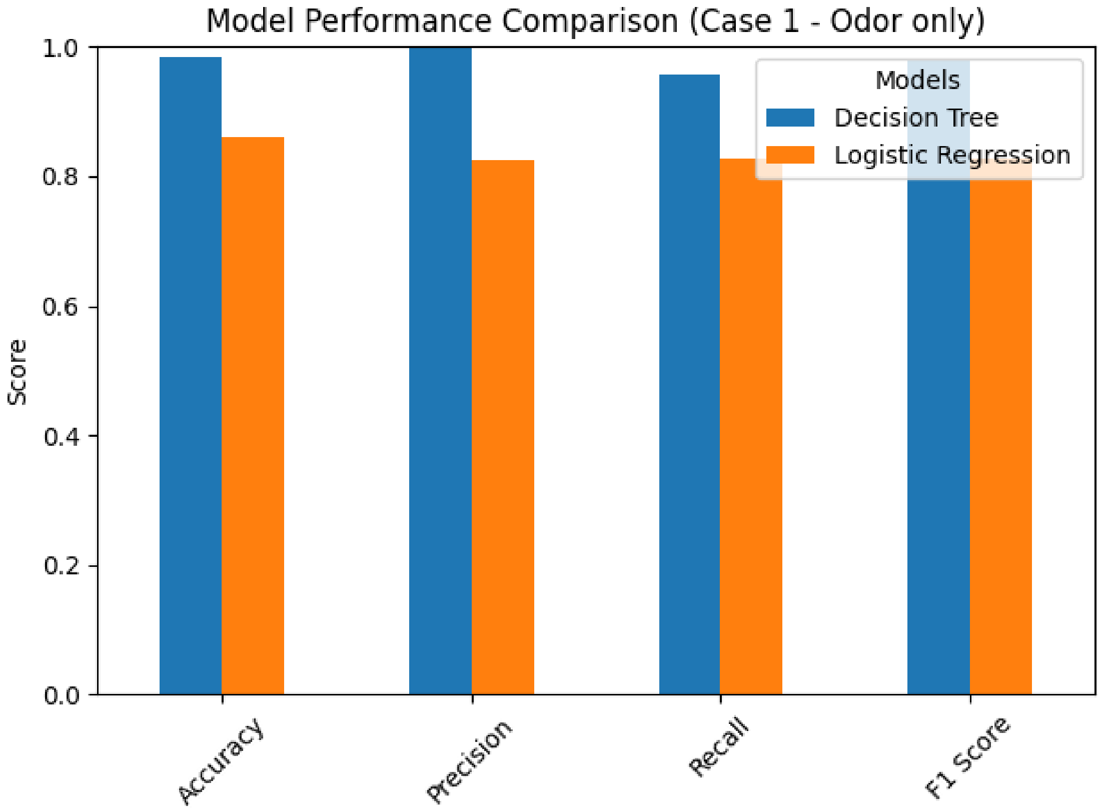
Taking these new models, let's put them side by side with their previous classifier to see how they performed.

5.2.1 Odor - Decision Tree vs Logistic Regression

```
Logistic Regression Model Performance (Case 1 - Odor only):
Accuracy: 0.8618
Precision: 0.8248
Recall: 0.8285
F1 Score: 0.8267
```

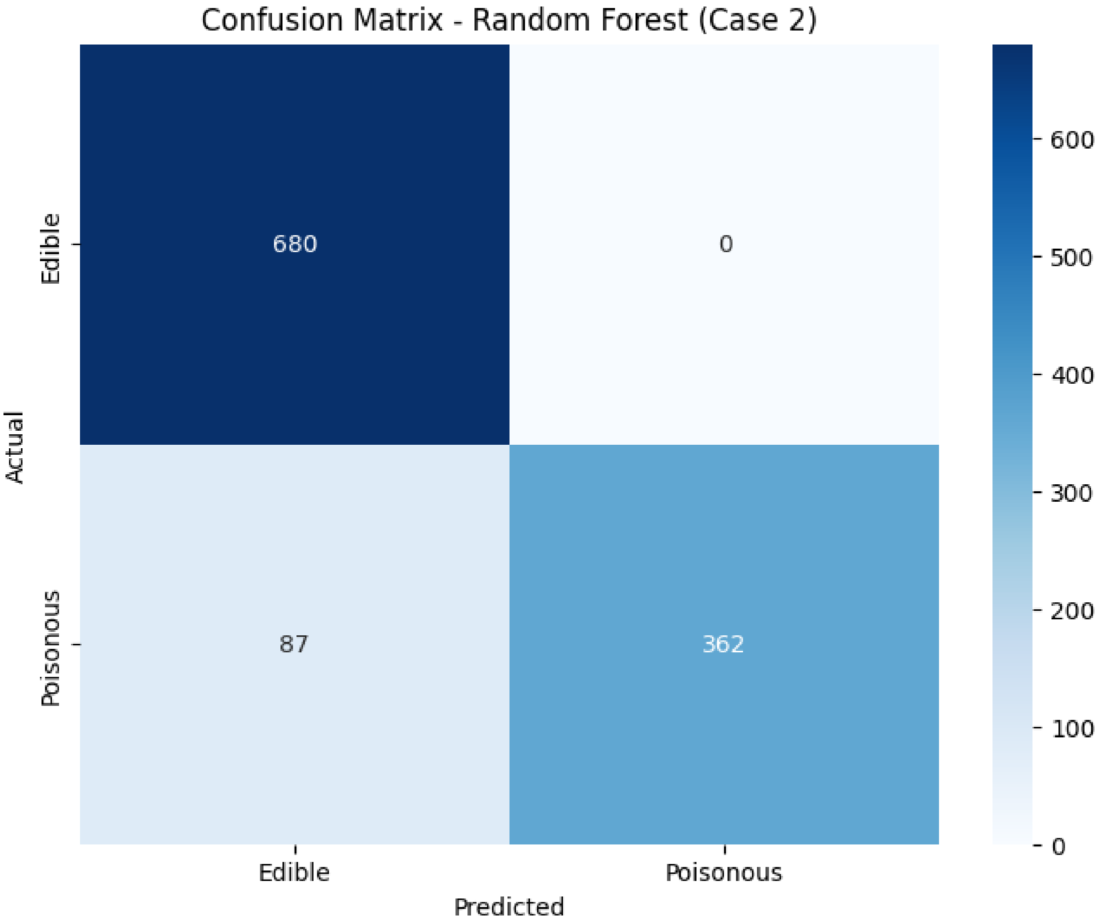

Confusion Matrix - Logistic Regression (Case 1)
Actual
Edible
Poisonous
601
79
77
372
Edible
Poisonous
Predicted
600
500
400
300
200
100

```
<Figure size 1000x600 with 0 Axes>
```

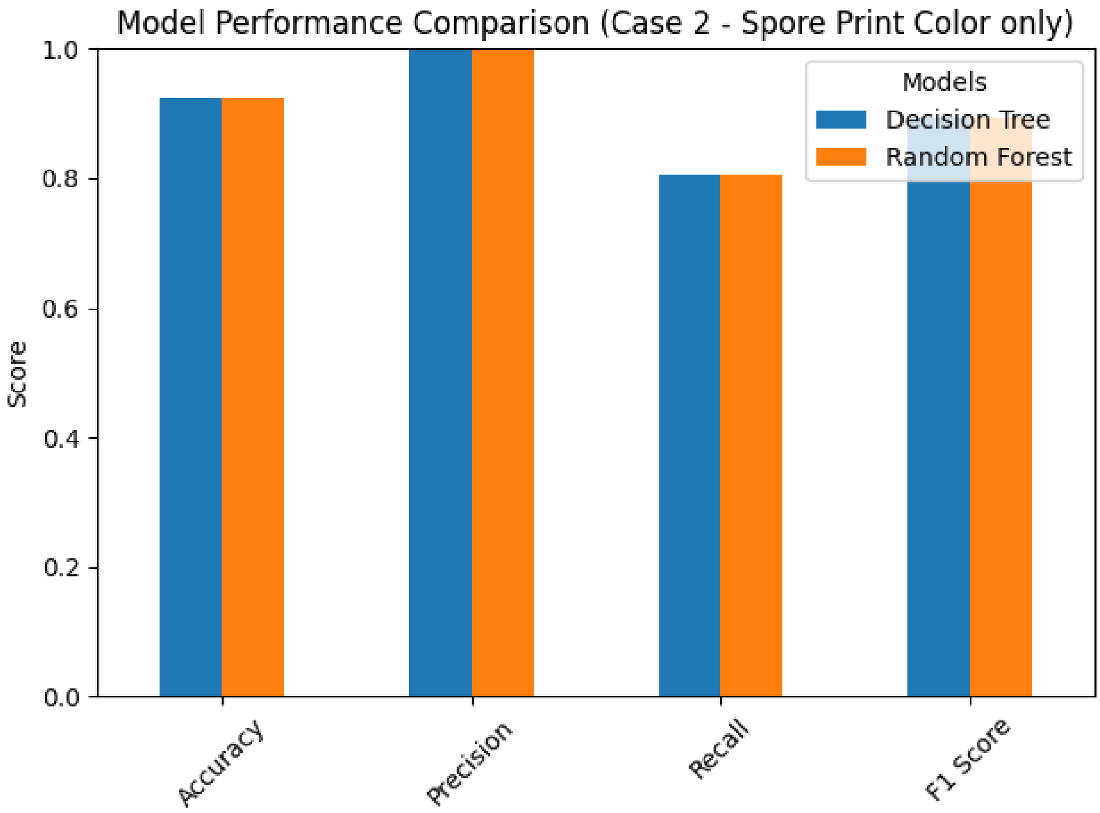


5.2.2 Spore Print Color - Decision Tree vs Random Forest

```
Random Forest vs Decision Tree prediction comparison:
Decision Tree predictions (first 10): [1 0 0 1 0 0 0 0 0 0]
Random Forest predictions (first 10): [1 0 0 1 0 0 0 0 0 0]
Predictions are IDENTICAL

Random Forest Model Performance (Case 2 - Spore Print Color only):
Accuracy: 0.9229
Precision: 1.0000
Recall: 0.8062
F1 Score: 0.8927
```

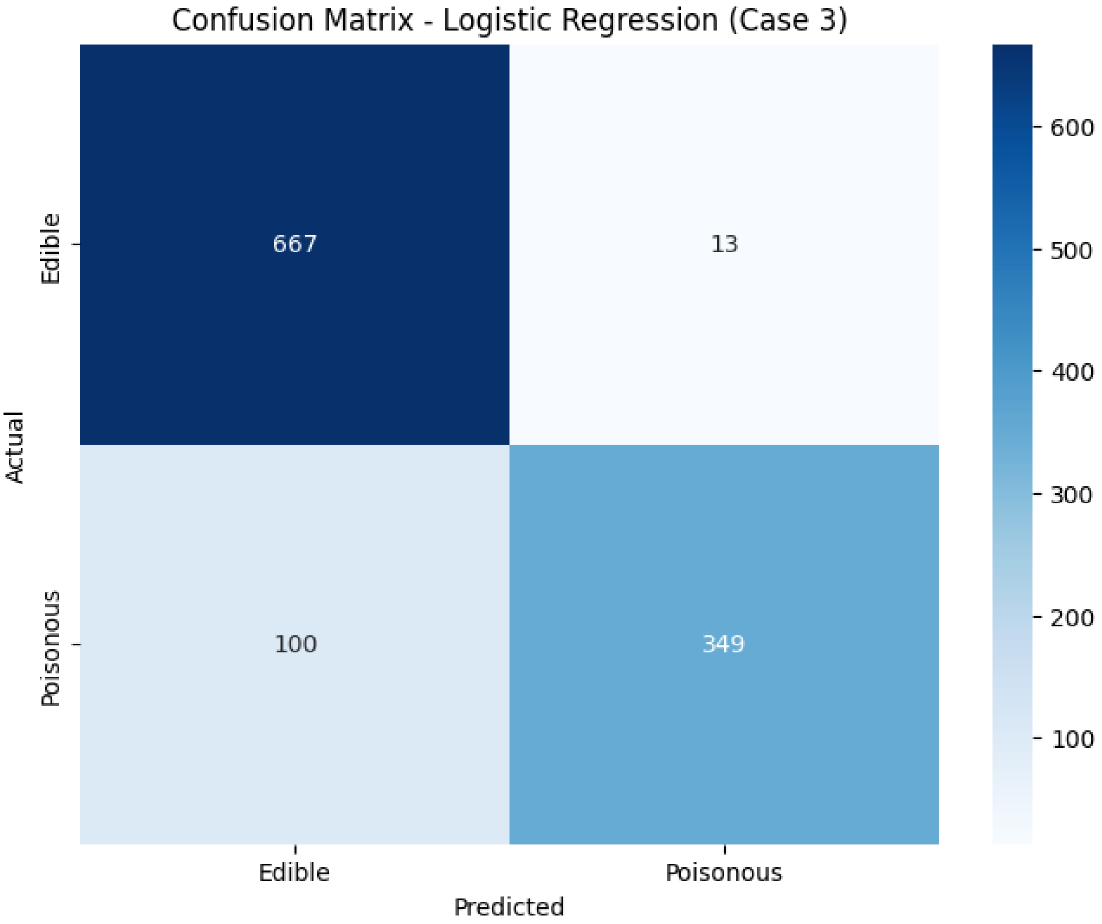

Confusion Matrix - Random Forest (Case 2)
Actual
Edible
Poisonous
680
0
87
362
Edible
Poisonous
Predicted
600
500
400
300
200
100
0

```
<Figure size 1000x600 with 0 Axes>
```

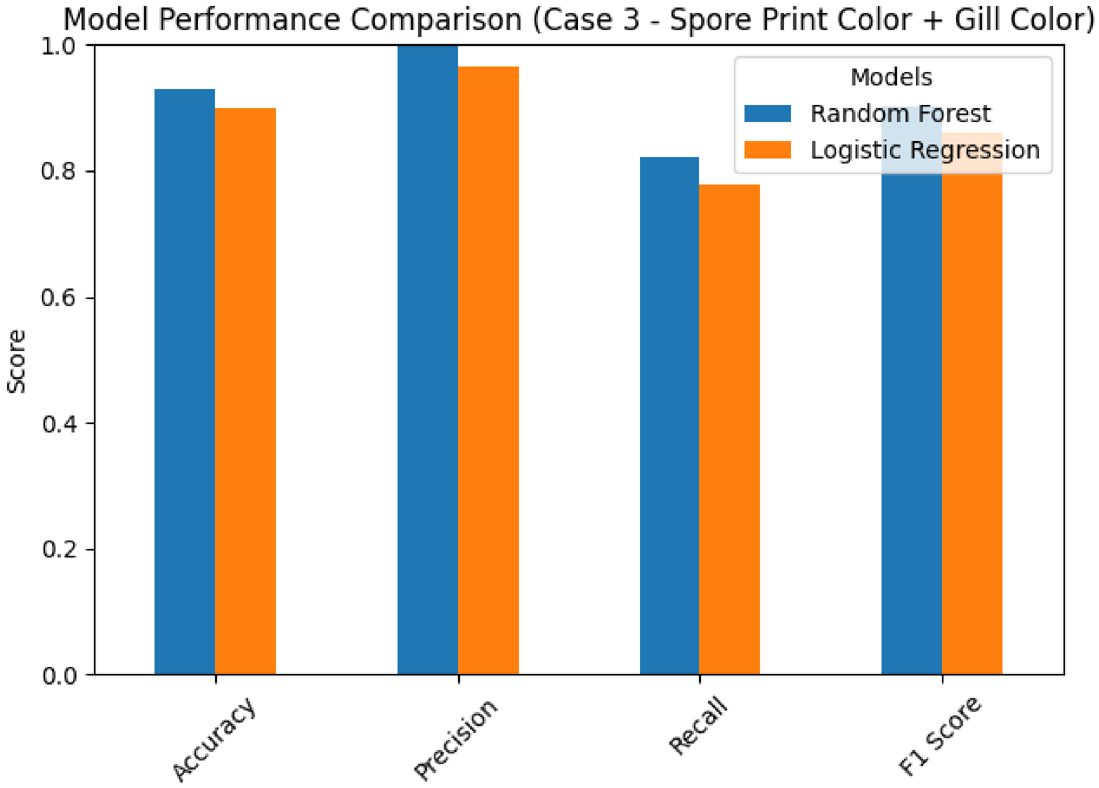


5.2.3 Spore Print Color + Gill Color - Random Forest vs Logistic Regression

```
Logistic Regression Model Performance (Case 3 - Spore Print Color + Gill Color):
Accuracy: 0.8999
Precision: 0.9641
Recall: 0.7773
F1 Score: 0.8607
```

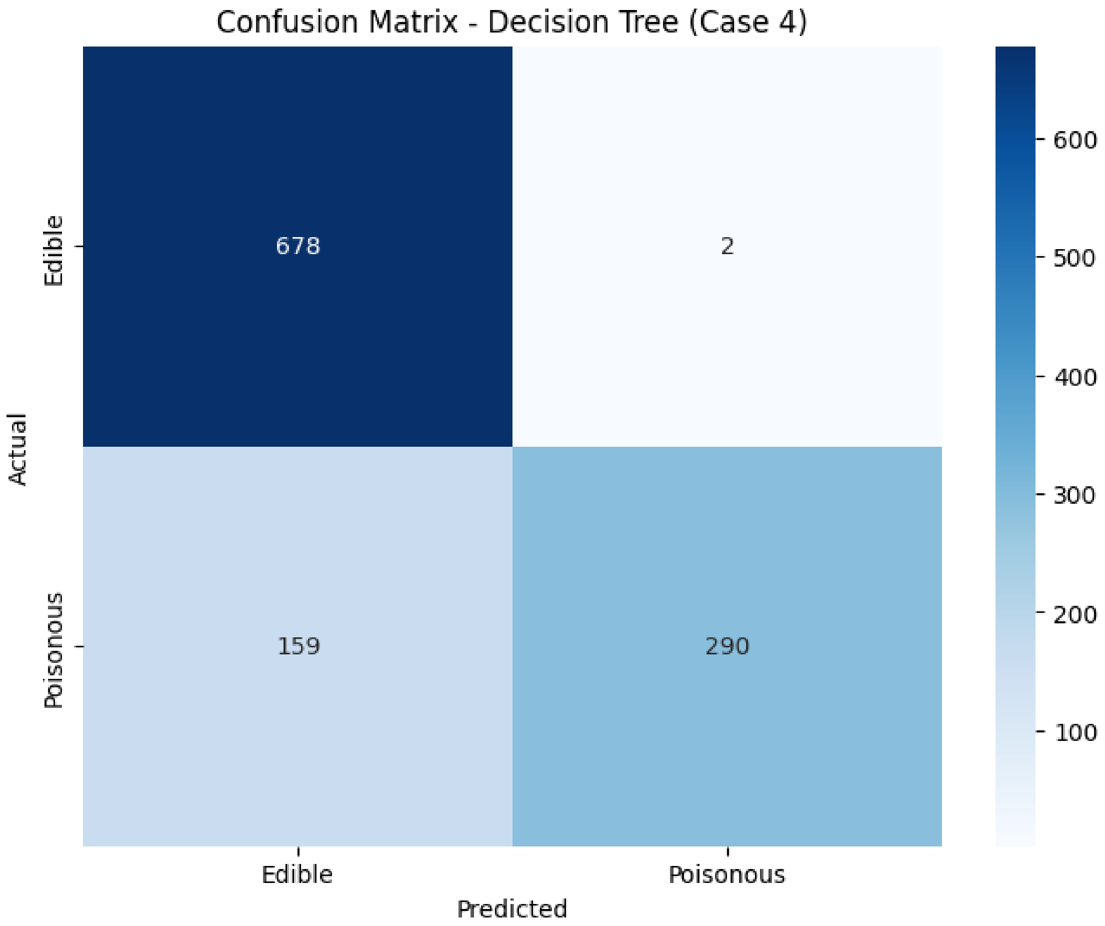

Confusion Matrix - Logistic Regression (Case 3)
Actual
Edible
Poisonous
667
13
100
349
Edible
Poisonous
Predicted
600
500
400
300
200
100

```
<Figure size 1000x600 with 0 Axes>
```

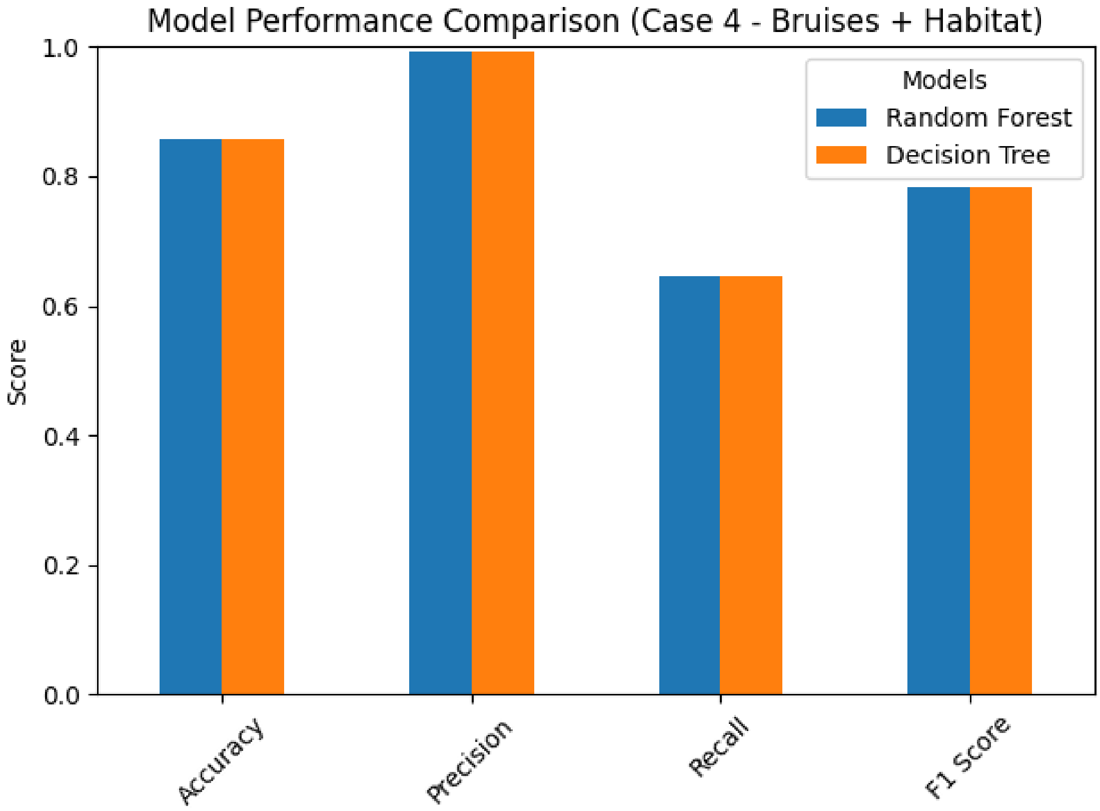


5.2.4 Bruises + Habitat - Random Forest vs Decision Tree Training

```
Random Forest vs Decision Tree prediction comparison:
Random Forest predictions (first 10): [0 0 0 1 0 0 0 0 0 0]
Decision Tree predictions (first 10): [0 0 0 1 0 0 0 0 0 0]
Predictions are IDENTICAL

Decision Tree Model Performance (Case 4 - Bruises + Habitat):
Accuracy: 0.8574
Precision: 0.9932
Recall: 0.6459
F1 Score: 0.7827
```


Confusion Matrix - Decision Tree (Case 4)
Actual
Edible
Poisonous
678
2
159
290
Edible
Poisonous
Predicted
600
500
400
300
200
100

```
<Figure size 1000x600 with 0 Axes>
```



Reflection 5: Which model performed better? Why might one classifier be more effective in this specific case?

- For odor (Case 1), Decision Tree outperformed Logistic Regression (98.32% vs 86.18%)
- For Cases 2 and 4, Decision Trees and Random Forests produced identical results. I examined the code in detail and everything looks to be intact. I am concluding this is due to:
  - Discriminative single features creating simple decision boundaries
  - Random Forests defaulting to decision-tree like splits when patterns are straightforward
- The Decision Tree's strong performance makes sense for this problem because:
  - The data has clear decision boundaries that align with Decision Tree's splitting approach
  - Mushroom classification likely follows rule-based patterns that trees naturally capture
  - The features have non-linear relationships that Decision Trees handle well

1.6 6 Conclusions

1.6.1 6.1 Summarize findings

- Certain mushroom characteristics are extremely predictive of edibility
- Odor alone can classify mushrooms with 98.32% accuracy. Having hunted them myself, I can confirm it's hard to beat a good schnoz in evaluating them.
- Models consistently avoided false negatives (never classified poisonous mushrooms as edible). Convenient given the high stakes of eating poison

- Decision Trees performed exceptionally well, suggesting the classification follows clear rules
- Simple models with few features achieved comparable performance to more complex approaches

1.6.2 6.2 Discuss challenges faced

- Missing values in the stalk-root feature required dropping data
- Identical performance between Decision Trees and Random Forests in some cases made comparison difficult
- Finding the optimal visual representations for categorical data with many unique values
- Determining the biological significance of the statistical patterns observed
- Getting models to produce different predictions for comparison was challenging

1.6.3 6.3 If you had more time, what would you try next?

- Given that we didn't use the data which was missing from the stalk column, I would be curious to insert values instead of drop these rows
- Previous projects have used Neural Networks - that would definitely be the next model I try!
- We are moving on to explore Ensemble Models in the course. I would be interesting to see how they work here
- There are plenty additional variables to test - would love to look deeper into some that don't have such strong correlation
- Most of all, I would like to create a general tool that can test models across CSVs. Specifically, one where you can select feature set(s), target variable, and classifiers and the code is generalized to run the rest. Good future project!

1.6.4 Reflection 6: What did you learn from this project?

- Some classification problems and associated datasets have inherently strong features that make even simple models effective
- Decision Trees work well when natural classification exists in the data
- Visualizing fundamentals about data before choosing feature sets can yield great results if the data and intuition are cooperating
- Sometimes simpler models with fewer features can outperform more complex approaches, as shown by our 1 feature prowess with Odor.