Mushroom Edibility Classification

Anieesh Saravanan and Pranav Akiri

“All mushrooms are edible; but some only once.”

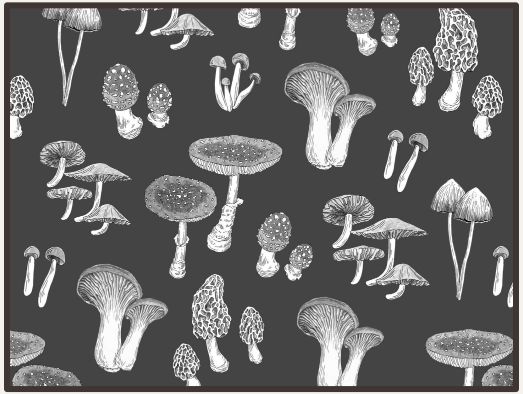
– Croatian Proverb

Table of contents

00
Introduction

Project Goal

- Develop a binary classification model for mushrooms
- Predict if mushrooms are edible or poisonous based on physical attributes
- Ensure public safety through accurate classification



Mushroom Overview

- Mushrooms are a section of a fungus
- Toxins used to prevent consumption
- **No** simple guidelines to identify mushrooms as poisonous
- Difference between edible and poisonous mushrooms are **extremely slight**
- Experienced mycologists make mistakes

Common Myths

Folklore has created many false truths about mushrooms:

Myth: Poisonous mushrooms always have bright and flashy colors

Myth: Snails, insects, or other animals won't eat poisonous mushrooms

Myth: Toxic mushrooms smell and taste horrible

Myth: Any mushroom becomes safe if cooked/processed enough

Impact

There are around **7,428** cases of exposure to toxic mushrooms, mostly by ingestion, each year.

01
Project Overview

Dataset Overview

- **Categorical** dataset with **8,124 instances** and **21 attributes**
- Class attribute is the edibility of the mushroom
- Sourced from **The Audubon Society Field Guide to North American Mushrooms**
- <u>Missing values</u>: 2,480 instances in the stalk-root attribute
- <u>Class distribution</u>: 52% edible, 48% poisonous (includes unknown classifications)

Attributes

class: edible = e, poisonous = p
cap-shape: bell = b, conical = c, convex = x, flat = f, knobbed = k, sunken = s
cap-surface: fibrous = f, grooves = g, scaly = y, smooth = s
cap-color: brown = n, buff = b, cinnamon = c, gray = g, green = r, pink = p, purple = u, red = e, white = w, yellow = y
bruises: true = t, false = f
odor: almond = a, anise = l, creosote = c, fishy = y, foul = f, musty = m, none = n, pungent = p, spicy = s
gill-attachment: attached = a, descending = d, free = f, notched = n
gill-spacing: close = c, crowded = w, distant = d
gill-size: broad = b, narrow = n
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
stalk-shape: enlarging = e, tapering = t
stalk-root: bulbous = b, club = c, cup = u, equal = e, rhizomorphs = z, rooted = r
stalk-surface-above-ring: fibrous = f, scaly = y, silky = k, smooth = s
stalk-surface-below-ring: fibrous = f, scaly = y, silky = k, smooth = s
stalk-color-above-ring: brown = n, buff = b, cinnamon = c, gray = g, orange = o, pink = p, red = e, white = w, yellow = y
stalk-color-below-ring: brown = n, buff = b, cinnamon = c, gray = g, orange = o, pink = p, red = e, white = w, yellow = y
veil-type: partial = p, universal = u
veil-color: brown = n, orange = o, white = w, yellow = y
ring-number: none = n, one = o, two = t
ring-type: cobwebby = c, evanescent = e, flaring = f, large = l, none = n, pendant = p, sheathing = s, zone = z
spore-print-color: black = k, brown = n, buff = b, chocolate = h, green = r, orange = o, purple = u, white = w, yellow = y
population: abundant = a, clustered = c, numerous = n, scattered = s, several = v, solitary = y
habitat: grasses = g, leaves = l, meadows = m, paths = p, urban = u, waste = w, woods = d

02 Preprocessing

Missing Value Handling

- Implemented the K-Nearest-Neighbors (KNN) algorithm
- Estimates missing values by analyzing the closest neighbors based on features
- Surpasses mode/deletion (reflects underlying patterns in the data)

```
knn_imputer = KNNImputer(n_neighbors = 5)
x_imputed = knn_imputer.fit_transform(x_encoded)
x = pd.DataFrame(x_imputed, columns = x.columns)
```

KNN Imputer implementation using scikit-learn method and depth 5 consideration

Feature Encoding

All of the data in the mushrooms dataset are categorical variables

Cap shape: bell (b) → 0, conical (c) → 1, flat (f) → 2

Odor: almond (a) → 0, fishy (y) → 8

```
for column in x.columns:

    le = LabelEncoder()
    x[column] = le.fit_transform(x[column].round().astype(int))
    label_encoders[column] = le
```

Encodings using scikit-learn functionality

Dataset Splitting

- Training Set: 80% of data for model training
- Validation Set: 10% for hyperparameter tuning
- Testing Set: 10% for unbiased evaluation of model performance

```
x_train, x_temp_split, y_train, y_temp_split = train_test_split(

    x_temp, y, test_size = 0.2, random_state = 42, stratify = y
)

x_val, x_test, y_val, y_test = train_test_split(

    x_temp_split, y_temp_split, test_size = 0.5, random_state = 42, stratify = y_temp_split

)
```

Split with stratified distribution using scikit-learn functionality

03 Manipulation

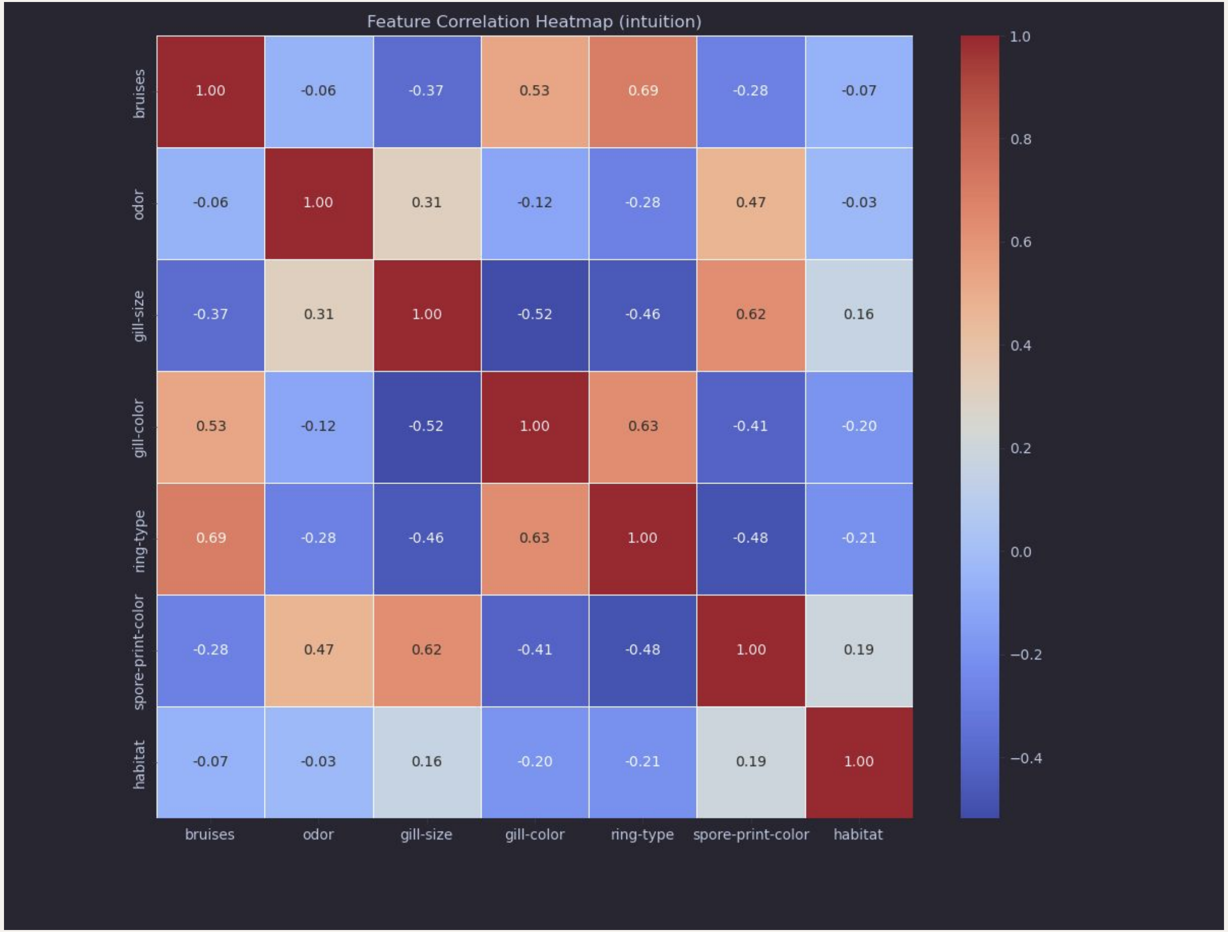
Attribute Selection Algorithms

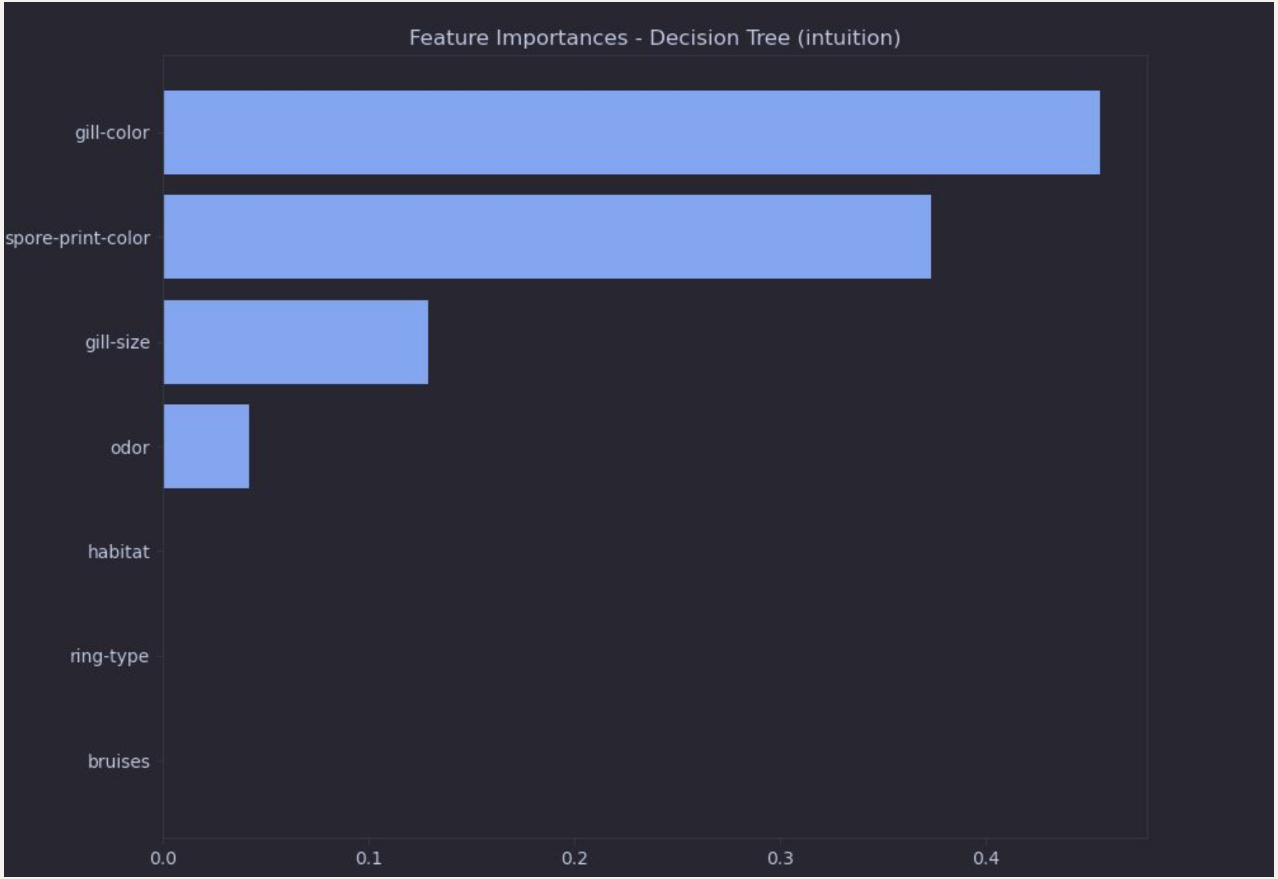
Selection Algorithm	Strengths	Weaknesses
Intuition	Simple to perform, takes the least amount of time	Prone to bias, can't be reproduced
Correlation Attribute Evaluation	Computationally efficient calculation in a straightforward way	Can only detect linear relationships, skewed by outliers
Gain Ratio Attribute Evaluation	Reduces overfitting, useful for Decision Trees	Ineffective with attributes that have only a few unique values
Information Gain Attribute Evaluation	Works with both categorical and discrete data	Biased toward attributes with many categories
Wrapper Subset Evaluation	Selection based on the specific model	Risks overfitting, especially when dataset is small

Intuition-Based Selection

- Basic selection based on personal consideration of the influence of attributes
- We watched National Geographic as kids (unofficial experts)

No.		
1	☐	bruises
2	☐	odor
3	☐	gill-size
4	☐	gill-color
5	☐	ring-type
6	☐	spore-print-color
7	☐	habitat
8	☐	class


Feature Correlation Heatmap (intuition)
bruises
odor
gill-size
gill-color
ring-type
spore-print-color
habitat
1.00
-0.06
-0.37
0.53
0.69
-0.28
-0.07
-0.06
1.00
0.31
-0.12
-0.28
0.47
-0.03
-0.37
0.31
1.00
-0.52
-0.46
0.62
0.16
0.53
-0.12
-0.52
1.00
0.63
-0.41
-0.20
0.69
-0.28
-0.46
0.63
1.00
-0.48
-0.21
-0.28
0.47
0.62
-0.41
-0.48
1.00
0.19
-0.07
-0.03
0.16
-0.20
-0.21
0.19
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Feature Importances - Decision Tree (intuition)
gill-color
spore-print-color
gill-size
odor
habitat
ring-type
bruises
0.0
0.1
0.2
0.3
0.4

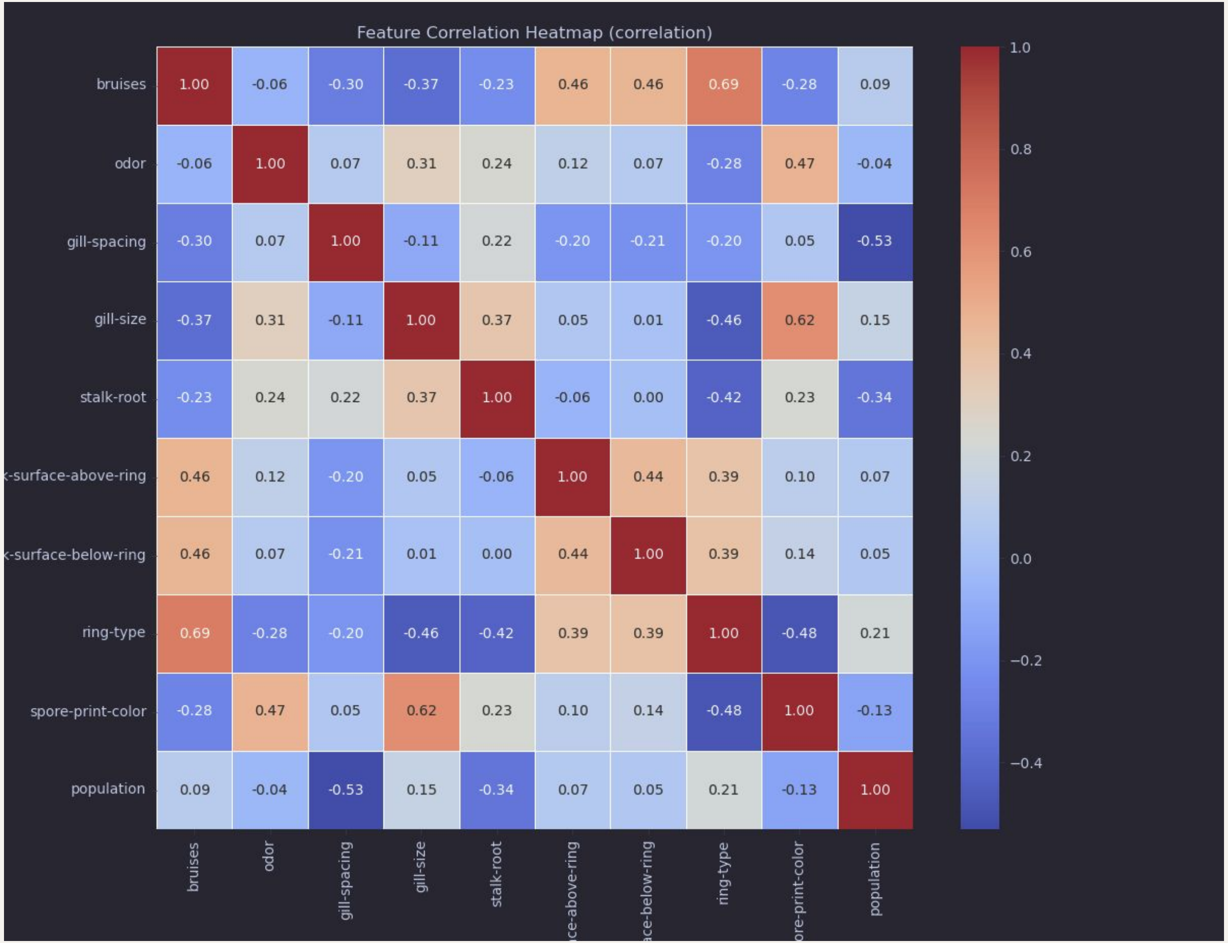
Correlation Attribute Evaluation

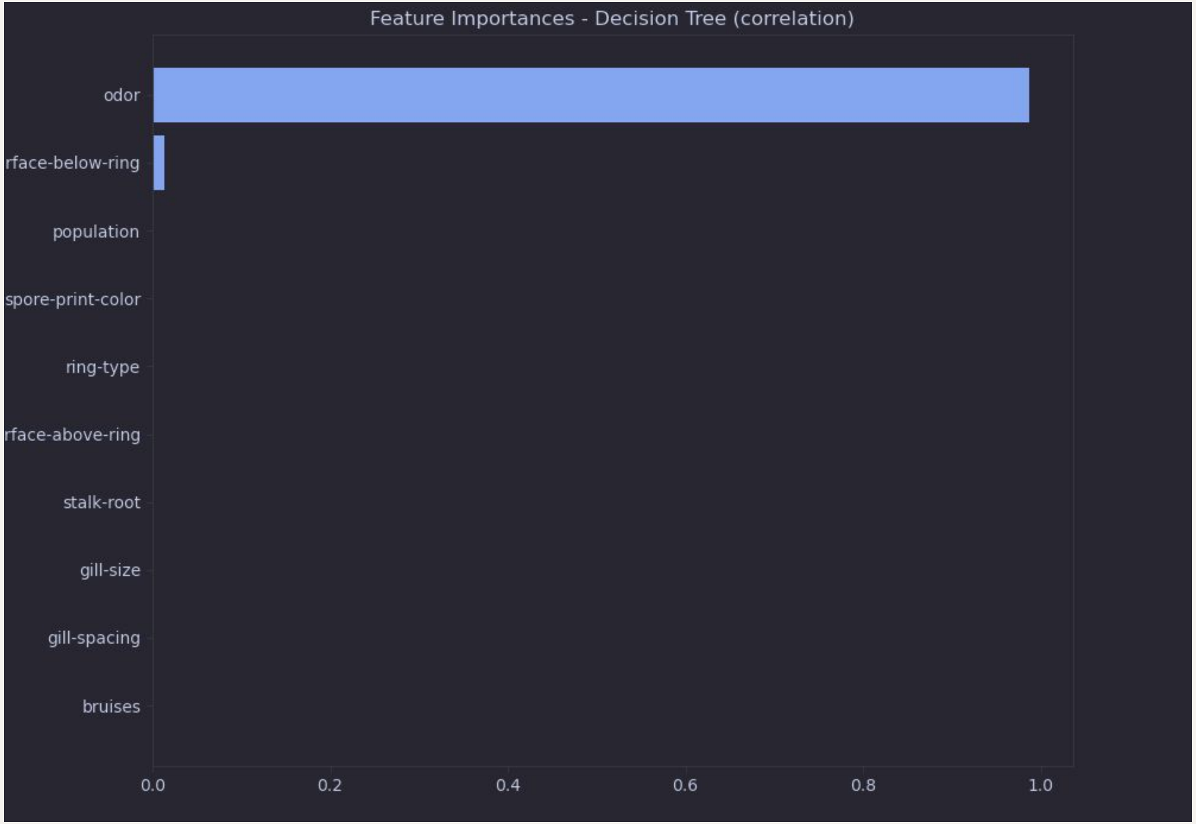
- Measures the linear correlation between attributes and class labels

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

```
Attribute Evaluator (supervised, Class (nominal): 23 class):
        Correlation Ranking Filter
Ranked attributes:
 0.5792    5 odor
 0.54      8 gill-size
 0.5015    4 bruises
 0.4928   12 stalk-surface-above-ring
 0.4341   13 stalk-surface-below-ring
 0.4131   19 ring-type
 0.3985   20 spore-print-color
 0.3484    7 gill-spacing
 0.3172   11 stalk-root
 0.2945   21 population
 0.242     9 gill-color
 0.2227   14 stalk-color-above-ring
 0.2187   15 stalk-color-below-ring
 0.1833   18 ring-number
 0.1675   22 habitat
 0.1396   17 veil-color
 0.1292    6 gill-attachment
 0.1213    2 cap-surface
 0.102    10 stalk-shape
 0.0753    3 cap-color
 0.0464    1 cap-shape
 0         16 veil-type

Selected attributes: 5,8,4,12,13,19,20,7,11,21,9,14,15,18,22,17,6,2,10,3,1,16 : 22
```

Cutoff value: 0.25


Feature Correlation Heatmap (correlation)
bruises
odor
gill-spacing
gill-size
stalk-root
ring-type
spore-print-color
population
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

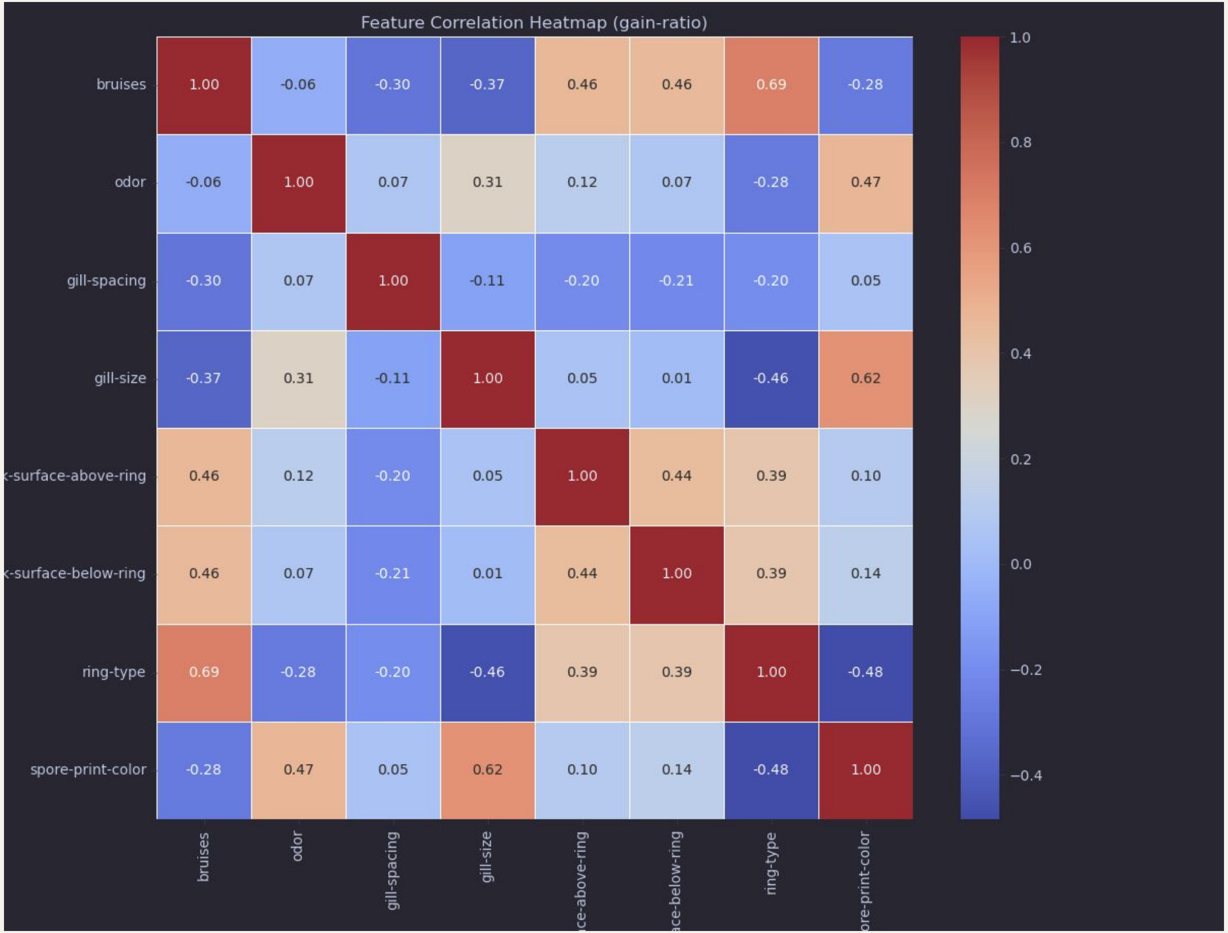

Feature Importances - Decision Tree (correlation)
odor
rface-below-ring
population
spore-print-color
ring-type
rface-above-ring
stalk-root
gill-size
gill-spacing
bruises
0.0
0.2
0.4
0.6
0.8
1.0

Gain Ratio Attribute Evaluation

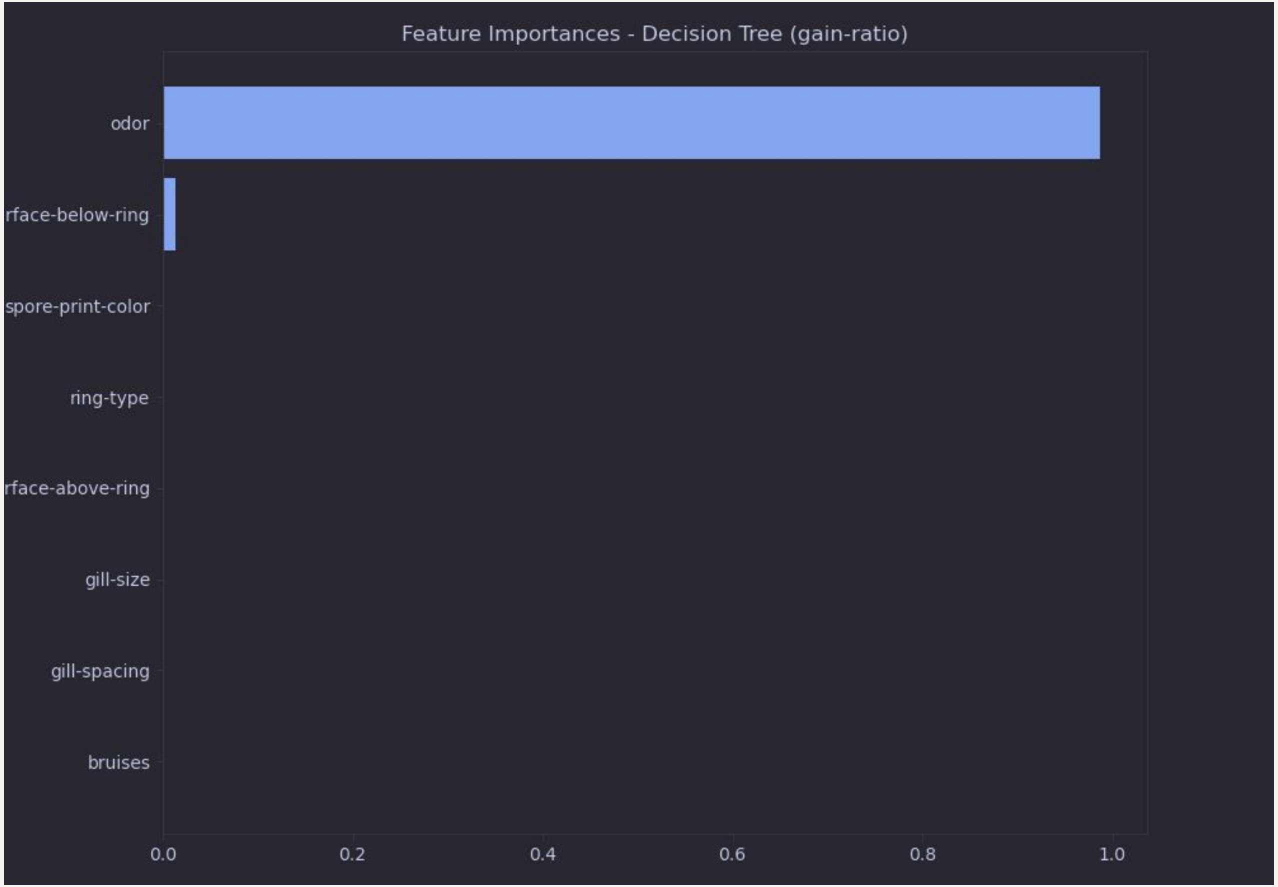
- Ranks attributes based on their gain ratio with respect to the class
- Gain ratio reduces bias towards attributes that have many distinct values

```
Attribute Evaluator (supervised, Class (nominal): 23 class):
        Gain Ratio feature evaluator

Ranked attributes:
 0.39065     5 odor
 0.25795     8 gill-size
 0.23312    12 stalk-surface-above-ring
 0.21818    20 spore-print-color
 0.20716    19 ring-type
 0.19644     4 bruises
 0.19433    13 stalk-surface-below-ring
 0.15815     7 gill-spacing
 0.1376      9 gill-color
 0.13106    14 stalk-color-above-ring
 0.12204    15 stalk-color-below-ring
 0.12137    17 veil-color
 0.10081    21 population
 0.10061    11 stalk-root
 0.09141    18 ring-number
 0.08182     6 gill-attachment
 0.06895    22 habitat
 0.02952     1 cap-shape
 0.01815     2 cap-surface
 0.01436     3 cap-color
 0.00762    10 stalk-shape
 0          16 veil-type

Selected attributes: 5,8,12,20,19,4,13,7,9,14,15,17,21,11,18,6,22,1,2,3,10,16 : 22
```

Cutoff value: 0.15



Feature Correlation Heatmap (gain-ratio)
bruises
odor
gill-spacing
gill-size
spore-print-color
ring-type
1.00
-0.06
-0.30
-0.37
0.46
0.46
0.69
-0.28
-0.06
1.00
0.07
0.31
0.12
0.07
-0.28
0.47
-0.30
0.07
1.00
-0.11
-0.20
-0.21
-0.20
0.05
-0.37
0.31
-0.11
1.00
0.05
0.01
-0.46
0.62
0.46
0.12
-0.20
0.05
1.00
0.44
0.39
0.10
0.46
0.07
-0.21
0.01
0.44
1.00
0.39
0.14
0.69
-0.28
-0.20
-0.46
0.39
0.39
1.00
-0.48
-0.28
0.47
0.05
0.62
0.10
0.14
-0.48
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



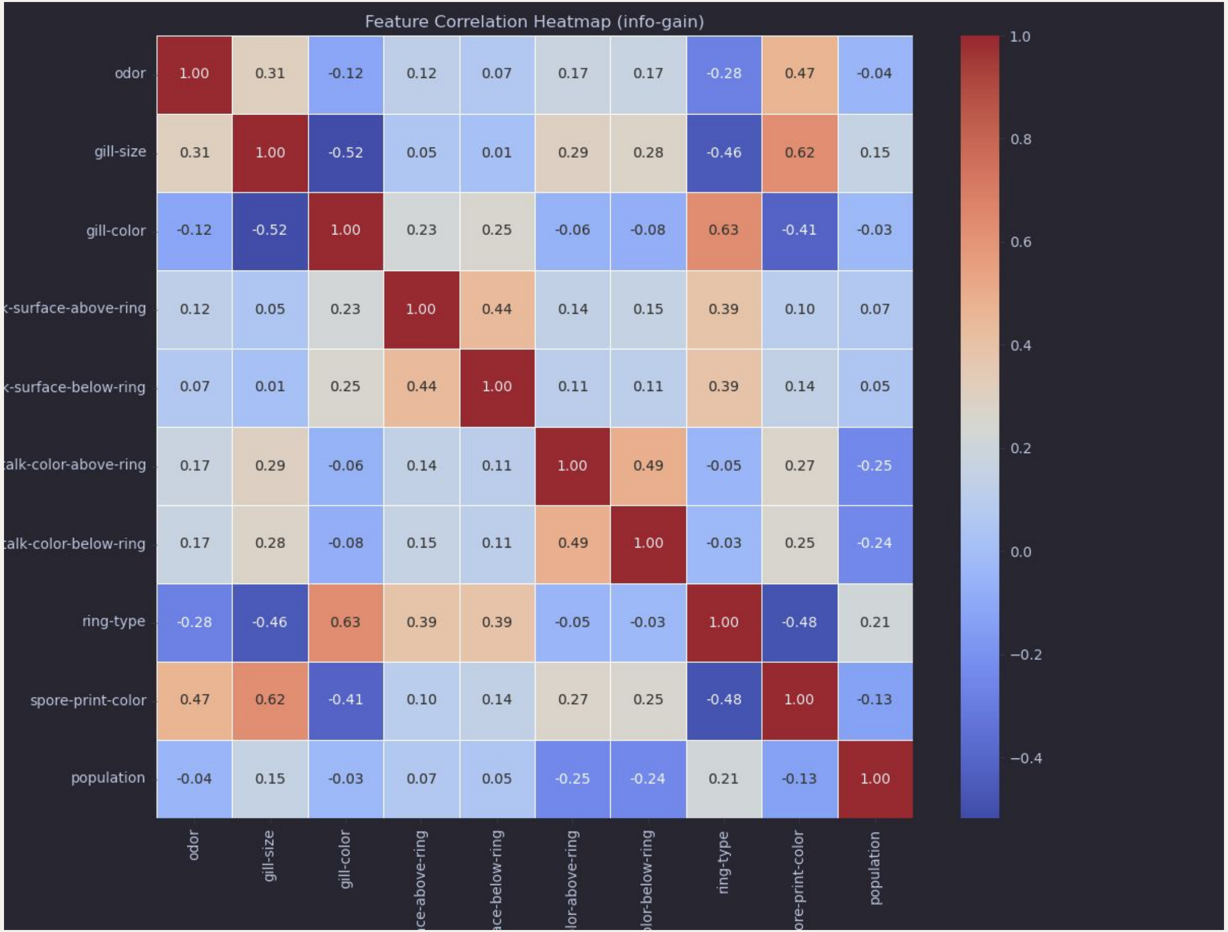
Feature Importances - Decision Tree (gain-ratio)
odor
rface-below-ring
spore-print-color
ring-type
rface-above-ring
gill-size
gill-spacing
bruises
0.0
0.2
0.4
0.6
0.8
1.0

Information Gain Attribute Evaluation

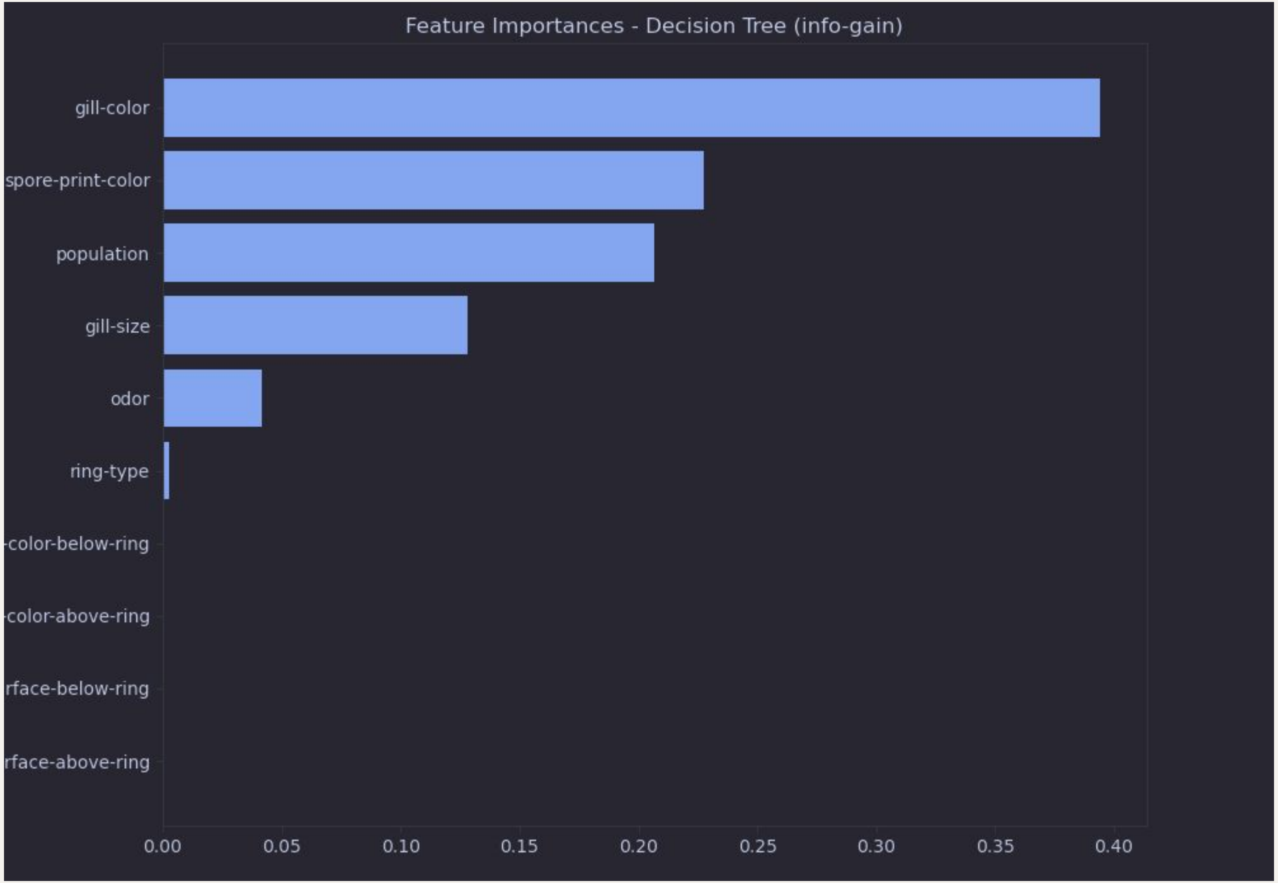
- Ranks attributes based on their information gain with respect to the class
- Information gain = reduction in entropy about the class

```
Attribute Evaluator (supervised, Class (nominal): 23 class):
        Information Gain Ranking Filter

Ranked attributes:
 0.90607     5 odor
 0.4807     20 spore-print-color
 0.41698     9 gill-color
 0.31802    19 ring-type
 0.28473    12 stalk-surface-above-ring
 0.27189    13 stalk-surface-below-ring
 0.25385    14 stalk-color-above-ring
 0.24142    15 stalk-color-below-ring
 0.23015     8 gill-size
 0.20196    21 population
 0.19238     4 bruises
 0.15683    22 habitat
 0.10835    11 stalk-root
 0.10088     7 gill-spacing
 0.0488      1 cap-shape
 0.03845    18 ring-number
 0.03605     3 cap-color
 0.02859     2 cap-surface
 0.02382    17 veil-color
 0.01417     6 gill-attachment
 0.00752    10 stalk-shape
 0          16 veil-type

Selected attributes: 5,20,9,19,12,13,14,15,8,21,4,22,11,7,1,18,3,2,17,6,10,16 : 22
```

Cutoff value: 0.2



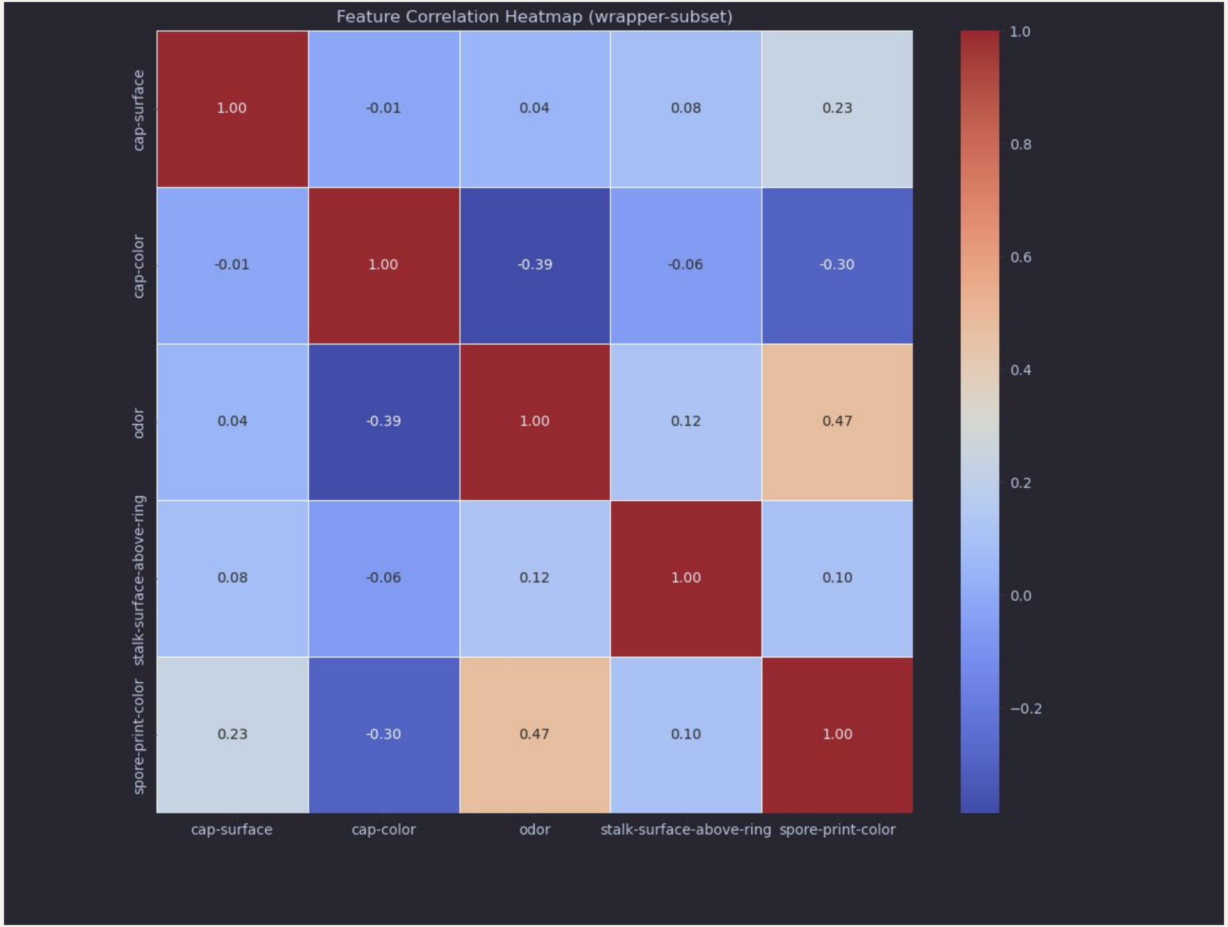
Feature Correlation Heatmap (info-gain)
odor
gill-size
gill-color
k-surface-above-ring
k-surface-below-ring
talk-color-above-ring
talk-color-below-ring
ring-type
spore-print-color
population

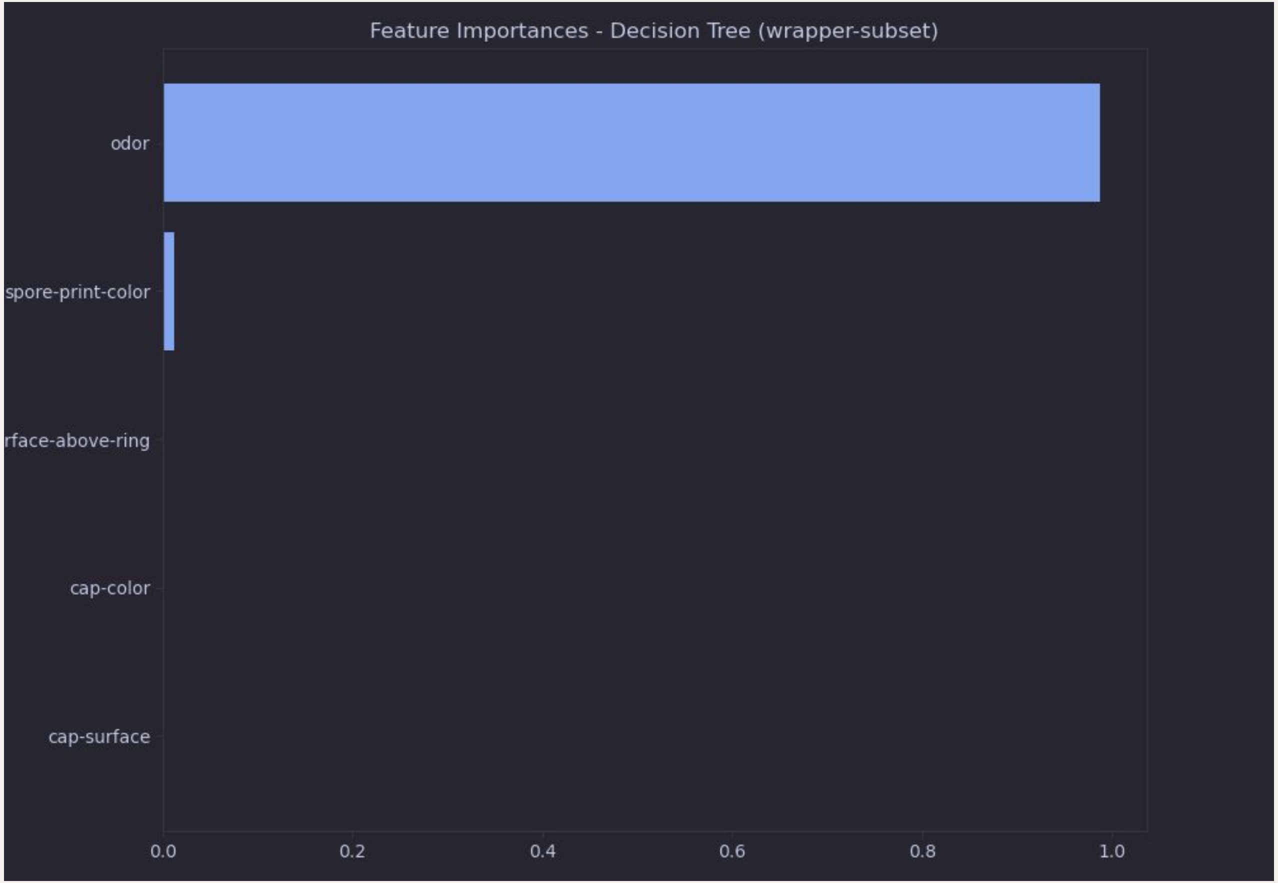

Feature Importances - Decision Tree (info-gain)
gill-color
spore-print-color
population
gill-size
odor
ring-type
-color-below-ring
-color-above-ring
rface-below-ring
rface-above-ring
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

Wrapper Subset Evaluation

- Evaluates the performance of a subset of attributes using J48 Decision Tree classifier
- Direct measure of the impact of selected attributes on model performance

```
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 182
        Merit of best subset found:    1

Attribute Subset Evaluator (supervised, Class (nominal): 23 class):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 2,3,5,12,20 : 5
                     cap-surface
                     cap-color
                     odor
                     stalk-surface-above-ring
                     spore-print-color
```


Feature Correlation Heatmap (wrapper-subset)
cap-surface
cap-color
odor
stalk-surface-above-ring
spore-print-color
1.00
-0.01
0.04
0.08
0.23
-0.01
1.00
-0.39
-0.06
-0.30
0.04
-0.39
1.00
0.12
0.47
0.08
-0.06
0.12
1.00
0.10
0.23
-0.30
0.47
0.10
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



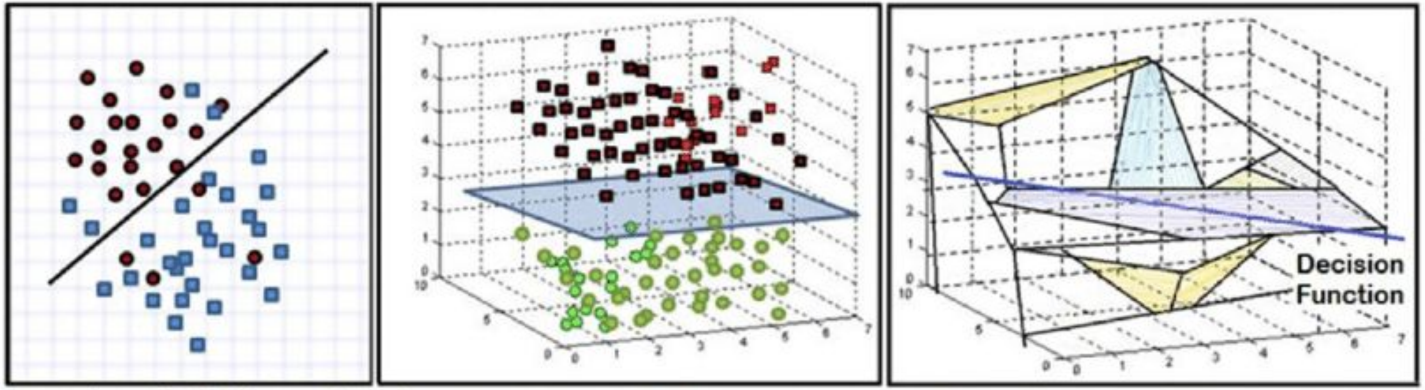
Feature Importances - Decision Tree (wrapper-subset)
odor
spore-print-color
rface-above-ring
cap-color
cap-surface
0.0
0.2
0.4
0.6
0.8
1.0

Classifier Models

- Decision Tree (J48): Uses rules to classify data by approximating a sine curve
- Quadratic Discriminant Analysis (QDA): Generates a quadratic boundary by fitting Gaussian densities to classes
- Logistic Regression: Utilizes a Bernoulli distribution to predicts probabilities for binary outcomes
- Support Vector Classifier (SVC): Finds optimal “hyperplane” for data (the decision boundary that maximizes the distance between the closest data points of two attributes)

Models were **trained using scikit-learn** and **stored with pickle** library serialization

Hyperplane



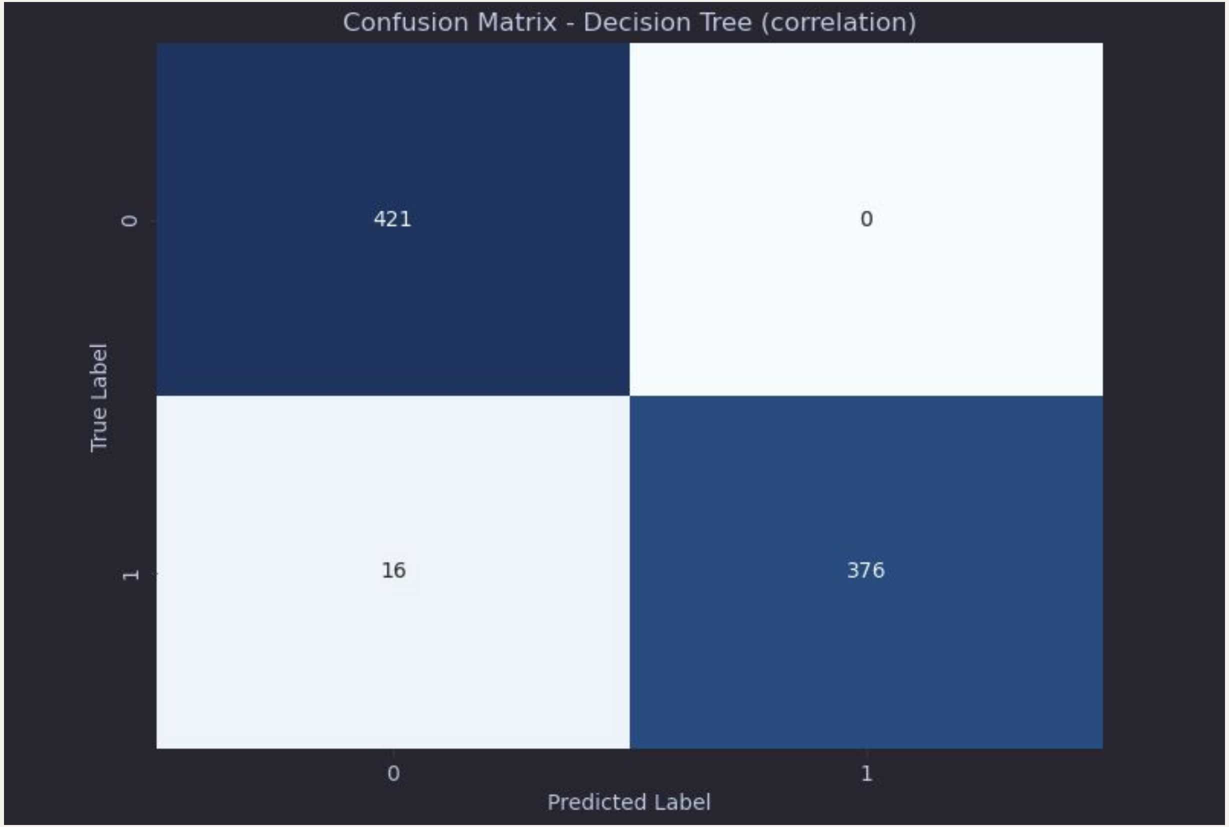
Hyperplane in 2-Dimensional Calculations (Line)

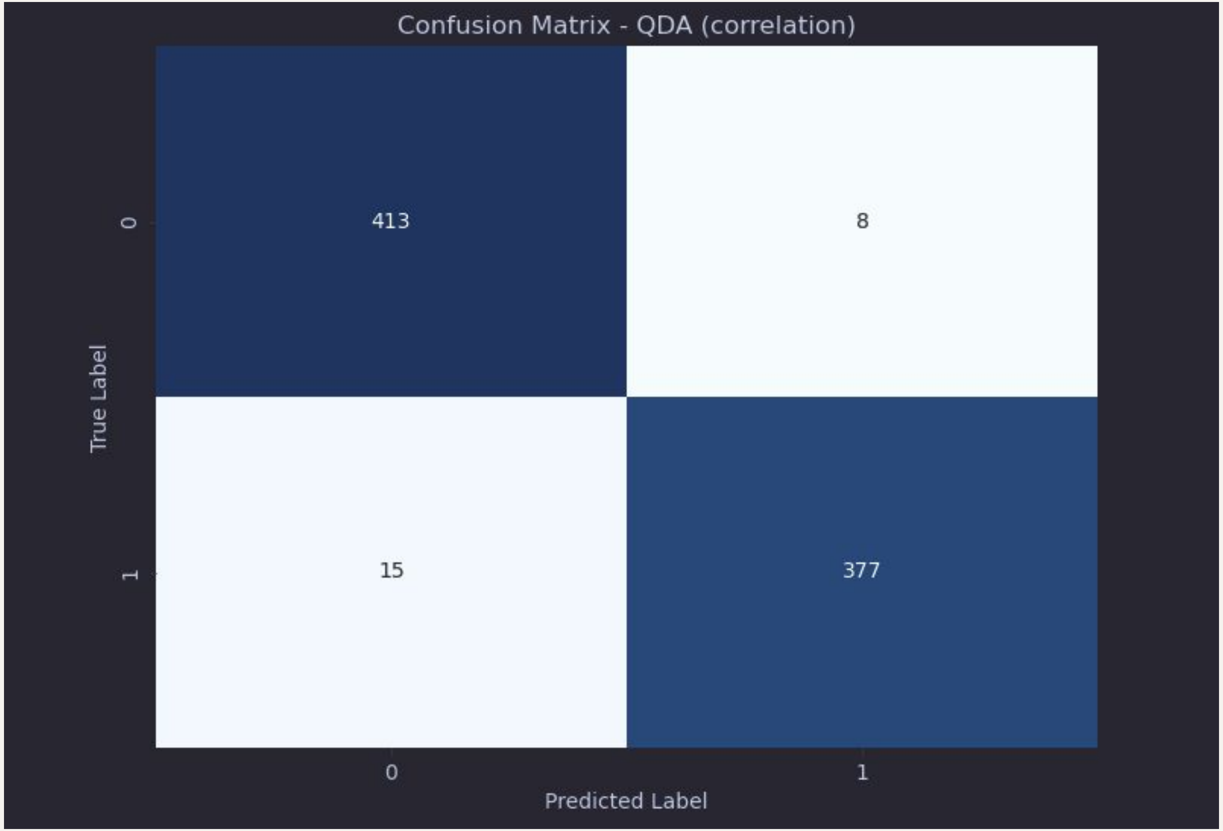
Hyperplane in 3- Dimensional Calculations (Plane)

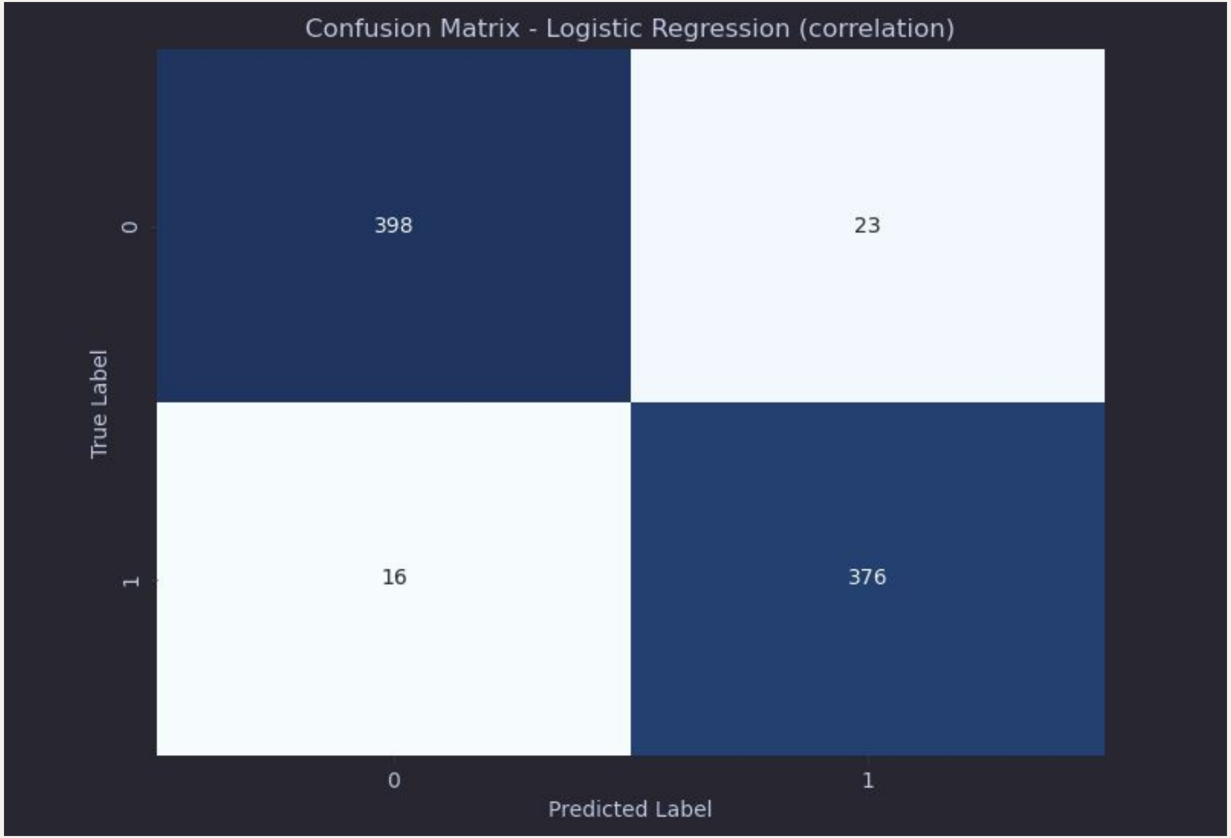
Hyperplane in n- Dimensional Calculations (Multiple Planes)

Comparison of Models

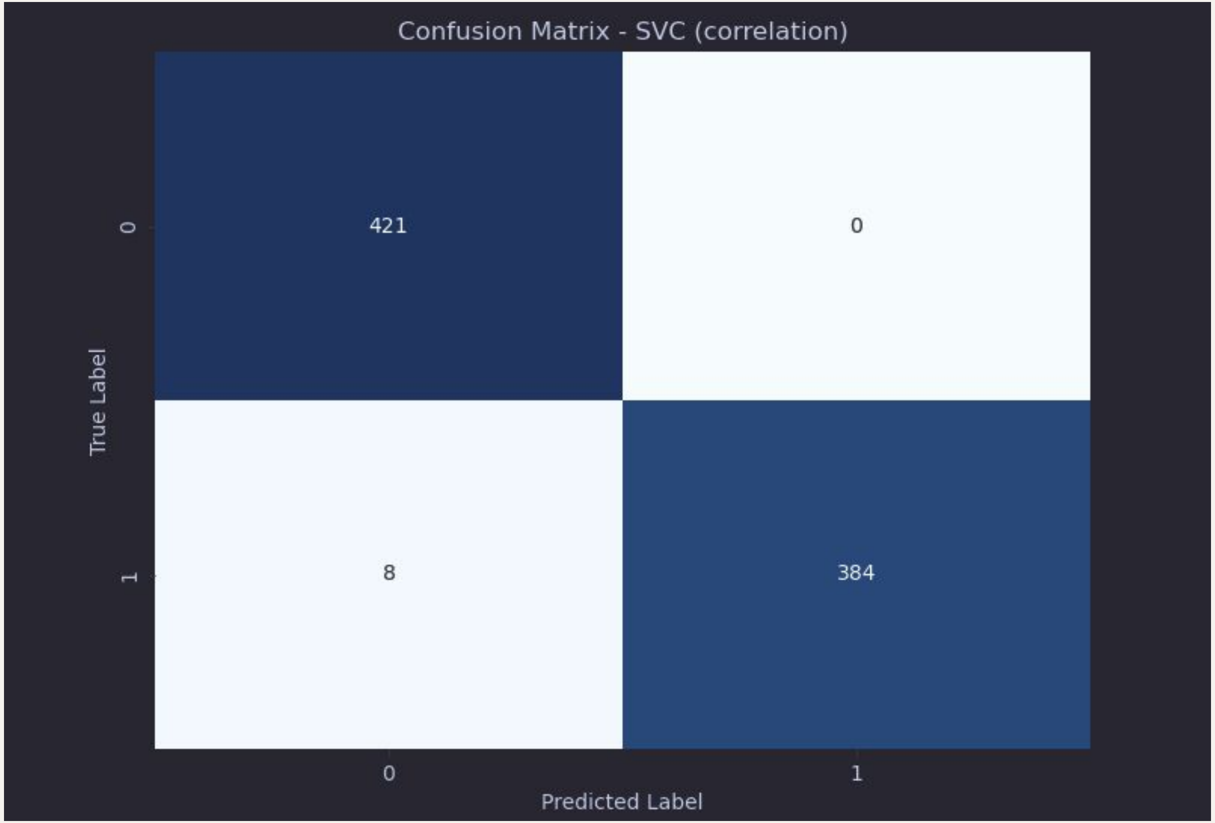
Model	Strengths	Weaknesses
J48 Decision Tree	Easy to use	Prone to overfitting
Logistic Regression	Good for binary classification	Struggles with non-linear relationships
SVC	Handles complex data well	Very slow
QDA	Works well with nonlinear relationships	Sensitive to dataset size


Confusion Matrix - Decision Tree (correlation)
True Label
0
1
421
0
16
376
0
1
Predicted Label


Confusion Matrix - QDA (correlation)
413
8
15
377
True Label
0
1
Predicted Label
0
1



Confusion Matrix - Logistic Regression (correlation)
True Label
0
1
398
23
16
376
0
1
Predicted Label



Confusion Matrix - SVC (correlation)
421
0
8
384
True Label
0
1
Predicted Label
0
1

04 Conclusion

Accuracy Results

	Decision Tree	QDA	Logistic Regression	SVC
Intuition Based Selection	94.9569%	93.6039%	88.9299%	98.2780%
CorrelationAttributeEval	98.0320%	97.1710%	95.2030%	99.0160%
GainRatioAttributeEval	98.0320%	94.7109%	94.7109%	98.8930%
InfoGainAttributeEval	95.4490%	96.0640%	89.7909%	97.0480%
WrapperSubsetEval	98.0320%	86.1009%	65.5597%	98.0320%

Support Vector Classifier (SVC) with CorrelationAttributeEval performed the best

ROC Area Results

	Decision Tree	*QDA*	*Logistic Regression*	*SVC*
Intuition Based Selection	0.980328	0.962359	0.924372	0.998764
CorrelationAttributeEval	0.990184	0.979822	0.980125	0.999727
GainRatioAttributeEval	0.990184	0.969218	0.957814	0.999952
InfoGainAttributeEval	0.979798	0.974835	0.947125	0.997225
WrapperSubsetEval	0.996801	0.938454	0.753908	0.999436

Support Vector Classifier (SVC) with GainRatioAttributeEval performed the best

Metrics for Chosen Model (SVC + Correlation)

```
Accuracy: 99.0160%
Correctly Classified Instances: 805
Incorrectly Classified Instances: 8
Kappa Statistic: 0.9803
Mean Absolute Error (MAE): 0.0098
Root Mean Squared Error (RMSE): 0.0992
Relative Absolute Error (RAE): 0.0197
Root Relative Squared Error (RRSE): 0.1985
Total Number of Instances: 813
               TP Rate   FP Rate  Precision    Recall  F-Measure        MCC  ROC Area  PRC Area
0             1.000000  0.018648   0.981352  1.000000   0.990588  0.980472  0.999727  0.999711
1             0.979592  0.000000   1.000000  0.979592   0.989691  0.980472  0.999727  0.999711
Weighted Avg  0.989796  0.009324   0.990343  0.990160   0.990155  0.980472  0.999727  0.999711

Confusion Matrix:
          Predicted 0   Predicted 1
Actual 0          421             0
Actual 1            8           384
```

Future Work

- Using Correlation as a selection algorithm could overlook attributes that might have high correlation when combined with other attributes
- Exploring alternative attribute selection algorithms
- Experimenting with different classifier models (e.g. Random Forest, Gradient Boosting)

Sources

Blythe, V. (n.d.). *Mushrooms* [Image]. Dribbble. https://dribbble.com/shots/14323122-Mushrooms

Brandenburg, W. E., & Ward, K. J. (2018, July 31). *Mushroom poisoning epidemiology in the United States*. National Library of Medicine. Retrieved October 23, 2024, from https://doi.org/10.1080/00275514.2018.1479561

Colors, C. (n.d.). *Glowing Mushrooms* [Image]. Dribbble. https://dribbble.com/shots/22519055-Glowing-Mushrooms

DecisionTreeClassifier. (n.d.). scikit-learn. Retrieved October 22, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

1.4. Support Vector Machines. (n.d.). scikit-learn. Retrieved October 22, 2024, from https://scikit-learn.org/1.5/modules/svm.html

Lillo, W. (n.d.). *Isometric Mushrooms* [Image]. Dribbble. https://dribbble.com/shots/20837558-Isometric-Mushrooms

LogisticRegression. (n.d.). scikit-learn. Retrieved October 22, 2024, from https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html

Sources

Miminoshvili, A. (n.d.). *Mushrooms* [Image]. Dribbble. https://dribbble.com/shots/3641736-Mushrooms

MKL47. (n.d.). *Mushrooms set* [Image]. Dribbble. https://dribbble.com/shots/23206484-Mushrooms-set

Package weka.attributeSelection. (n.d.). Weka Documentation. Retrieved October 22, 2024, from https://weka.sourceforge.io/doc.dev/weka/attributeSelection/package-summary.html

QuadraticDiscriminantAnalysis. (n.d.). scikit-learn. Retrieved October 22, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis.html

Schultz, P. (n.d.). *Mushroom Pattern* [Image]. Dribbble. https://dribbble.com/shots/5709006-Mushroom-Pattern

Wild Mushrooms. (n.d.). The Blue Ridge Poison Center. Retrieved October 23, 2024, from https://med.virginia.edu/brpc/wp-content/uploads/sites/274/2015/10/Mushrooms.pdf

Thank you!