Mushroom Edibility Classification

Sergio Robledo

Project Overview

- Purpose
  - Make mushroom foraging beginner friendly & safe
- Desired Outcome
  - Machine learning model to predict mushroom edibility
    - Achieve 100% accuracy on test set
- Project Deliverables
  - Milestone Report 1 (Exploration & Visualizations)
  - Milestone Report 2 (Machine Learning Model)
  - Final Report
  - Presentation Slides
  - Project Code (with trained classifier)

Mushroom Dataset & Preprocessing

Mushroom Dataset Summary

- Dataset Overview
  - Mushroom Classes:
    - Edible or Poisonous
    - 8,124 Mushrooms Observed
  - Features:
    - Number: 22
    - Type: Nominal
    - Examples: Odor, Cap Color, Habitat, etc.
- Source: UCI Machine Learning Repository

Data Preprocessing

	class	cap-shape	cap-surface	cap-color	bruises	odor
0	p	x	s	n	t	p
1	e	x	s	y	t	a
2	e	b	s	w	t	l
3	p	x	y	w	t	p
4	e	x	s	g	f	n

	edible	cap_shape	cap_surface	cap_color	bruises	odor
0	False	convex	smooth	brown	True	pungent
1	True	convex	smooth	yellow	True	almond
2	True	bell	smooth	white	True	anise
3	False	convex	scaly	white	True	pungent
4	True	convex	smooth	gray	False	none

1. Target Column
    a. Rename from “class” to “edible”
    b. Replace categorical “e” and “p” values with booleans
        i. True for Edible
2. Rename Feature Columns
    a. Replace hyphens (“-”) with underscores (“_”)
3. Map single-letter feature values to their full name

Exploration & Visualization

Class Counts & Edibility Correlations

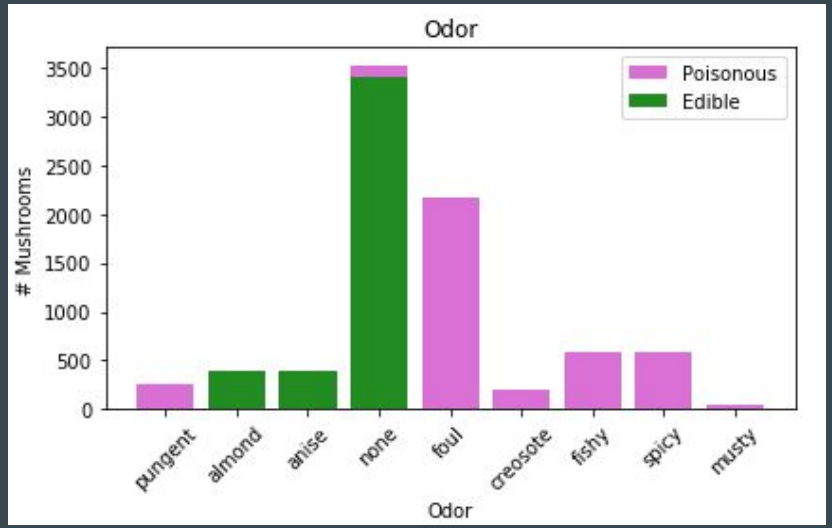
- ➔ Class Counts
  - ◆ Fairly Balanced
    - Classes = 50%
- ➔ Edibility vs. Feature Correlations
  - ◆ Pearson Correlation Coefficient
  - ◆ Strongest Correlations
    - Positive (Most Likely Edible)
      - No Odor
    - Negative (Most Likely Poisonous)
      - Foul Odor
  - ◆ No Correlation
    - Veil Type “Partial”
      - All mushrooms observed had a “partial” veil.

```
Total # of Observed Mushrooms: 8,124

                        Edible: 4,208 (52)%
          Poisonous Mushrooms: 3,916 (48)%
```

```
Mushroom Edibility Correlations:
--------------------------------

odor_none                          0.785557
ring_type_pendant                  0.540469
gill_size_broad                    0.540024
bruises_True                       0.501530
stalk_surface_above_ring_smooth    0.491314
                                       ...
gill_size_narrow                  -0.540024
stalk_surface_below_ring_silky    -0.573524
stalk_surface_above_ring_silky    -0.587658
odor_foul                         -0.623842
veil_type_partial                       NaN
```

Odor

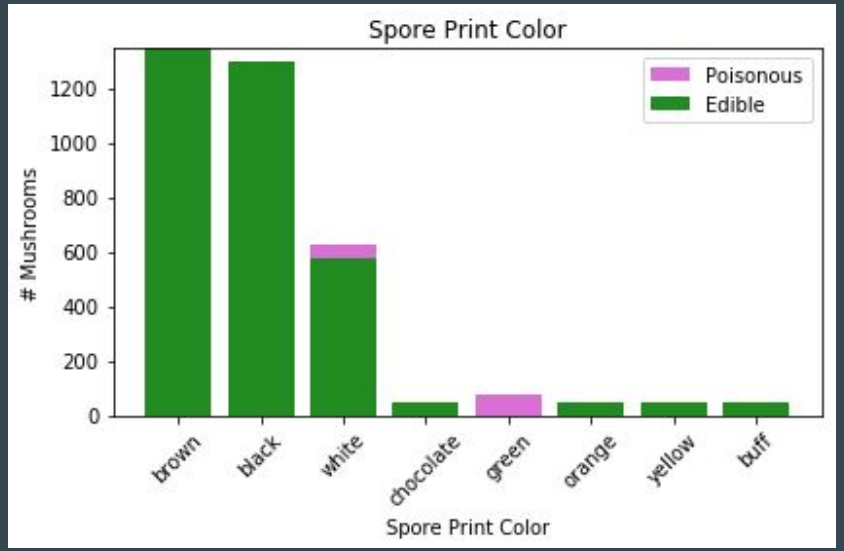
- ➔ Total No. of Mushrooms: 8,124
  - ◆ Edible: 4,208
  - ◆ Poisonous: 3,916
- ➔ Why Odor?
  - ◆ Strongest Initial Correlation
  - ◆ Almost Perfect Separation
  - ◆ Misunderstood Poisonous Mushrooms:
    - Reduced from 3,916 to 120
- ➔ Remaining Mixed Categories
  - ◆ No Odor



1. Mushroom Data
   1.1. All Odors

Spore Print Color

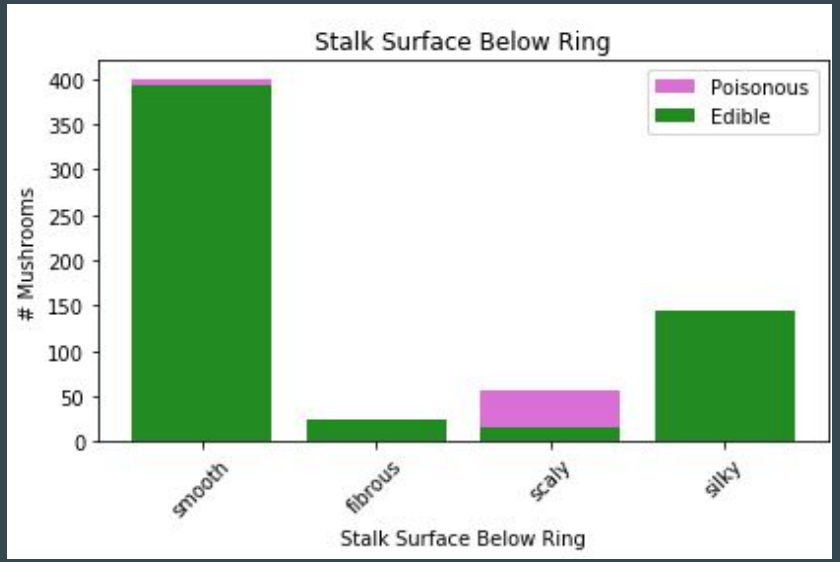
- ➔ Total No. of Mushrooms: 3,528
  - ◆ Edible: 3,408
  - ◆ Poisonous: 120
- ➔ Why Spore Print Color?
  - ◆ After Filtering Dataset
    - Strongest Edibility Correlation
  - ◆ No Odor & Green Spore Prints
    - All Poisonous (72)
- ➔ Remaining Mixed Categories
  - ◆ White Spore Print



1. Mushroom Data
   1.1. No Odor

Stalk Surface Below Ring

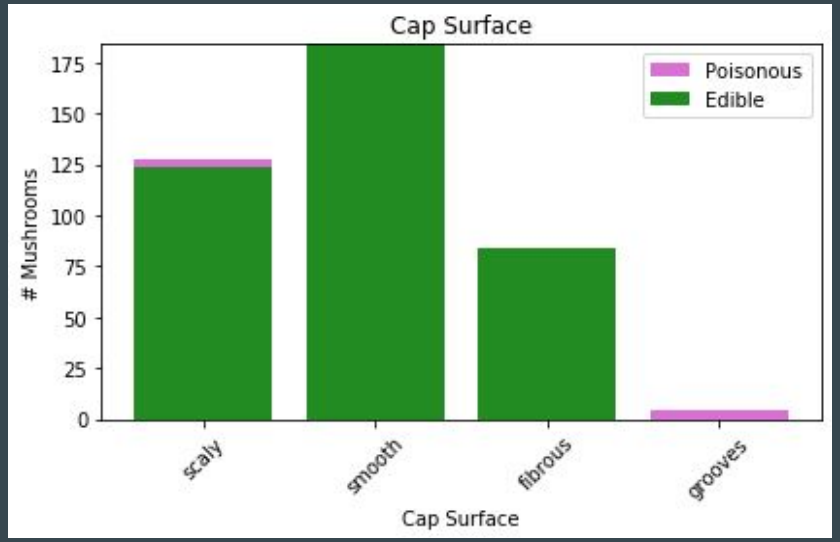
- Total No. of Mushrooms: 624
  - Edible: 576
  - Poisonous: 48
- Remaining Mixed Mushrooms
  - Smooth Stalk Below Ring
  - Scaly Stalk Below Ring



1. Mushroom Data
   1.1. No Odor
   1.2. White Spore Print

Cap Surface (Smooth Stalk Below Ring)

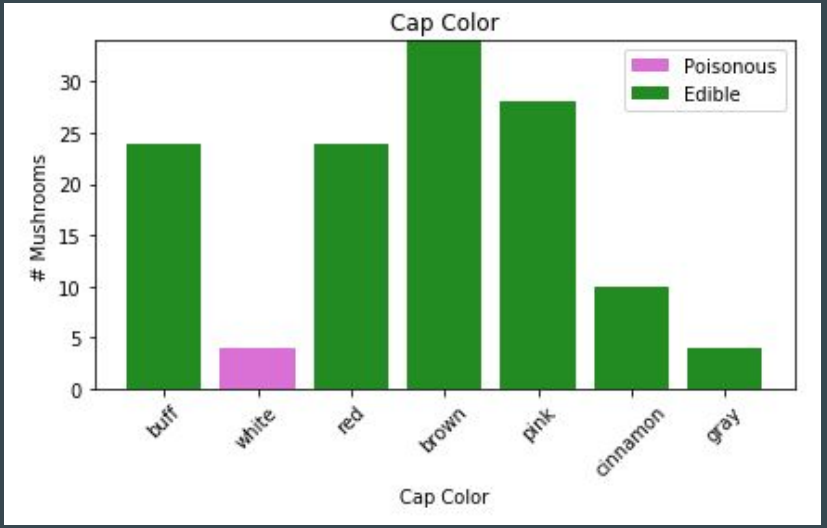
- ➔ Total No. of Mushrooms: 400
  - ◆ Edible: 392
  - ◆ Poisonous: 8
- ➔ Groovy Caps = Poisonous
  - ◆ 4 Mushrooms
- ➔ Remaining Mixed Mushrooms
  - ◆ Scaly Cap



1. Mushroom Data
   1.1. No Odor
   1.2. White Spore Print
   1.3. Smooth Stalk Below Ring

Cap Color (Smooth Stalk Below Ring)

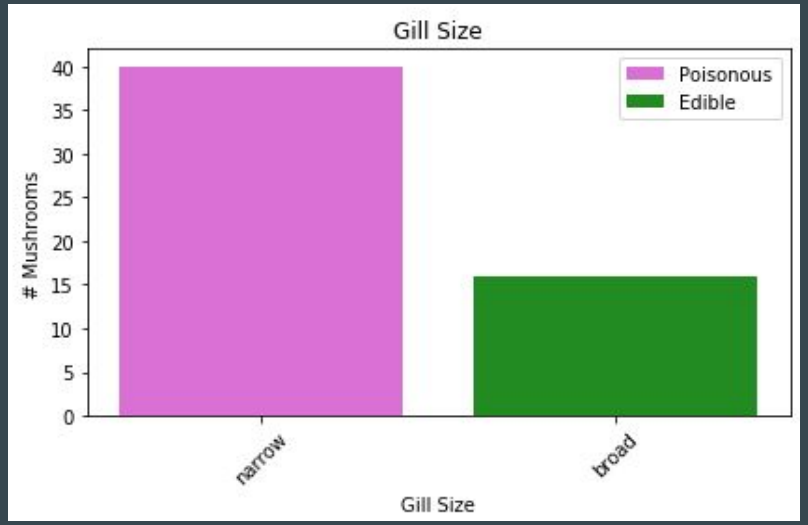
- Total No. of Mushrooms: 128
  - Edible: 124
  - Poisonous: 4
- White Caps = Poisonous
  - Final 4 Poisonous Mushrooms
- Most Important Features
  - Odor
  - Spore Print Color
  - Stalk Surface Below Ring
  - Cap Surface
  - Cap Color



1. Mushroom Data
   1.1. No Odor
   1.2. White Spore Print
   1.3. Smooth Stalk Below Ring
   1.4. Scaly Cap

Gill Size (Scaly Stalk Below Ring)

- ➔ Total No. of Mushrooms: 56
  - ◆ Edible: 16
  - ◆ Poisonous: 40
- ➔ Narrow Gills = Poisonous
  - ◆ Last Needed Feature
  - ◆ 40 Poisonous Mushrooms
- ➔ Most Important Features
  - ◆ Odor
  - ◆ Spore Print Color
  - ◆ Stalk Surface Below Ring
  - ◆ Gill Size



1. Mushroom Data
   1.1. No Odor
   1.2. White Spore Print
   1.3. Scaly Stalk Below Ring

Machine Learning

One-Hot Encoding

	odor
0	pungent
1	almond
2	anise
3	pungent
4	none

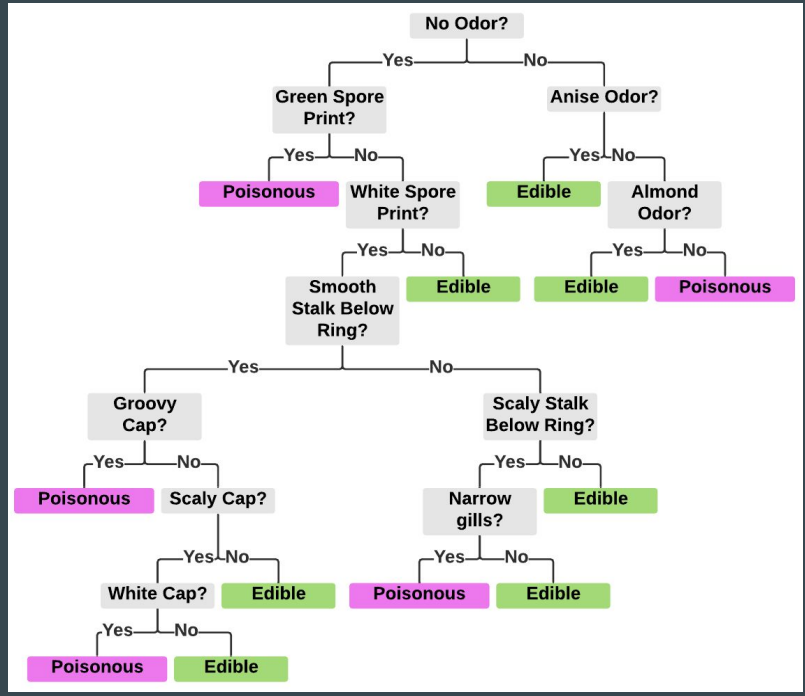
	odor_anise	odor_creosote	odor_fishy	odor_foul	odor_musty	odor_none	odor_pungent	odor_spicy
0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0
4	0	0	0	0	0	1	0	0

- ➔ One-Hot Encoding
  - ◆ Models accept integers & floats
    - Not strings
  - ◆ Every Category = 1 Column
    - Ex: Odor Feature = 5 Columns
      - Drop one redundant column
    - 6 Features = 32 Columns

- ➔ Encoded Features
  - ◆ Odor
  - ◆ Spore Print Color
  - ◆ Stalk Surface Below Ring
  - ◆ Cap Surface
  - ◆ Cap Color
  - ◆ Gill Size

Optimal Decision Tree

- ➔ Nominal Data
  - ◆ Decision Trees Handle Very Well
- ➔ Optimal Decision Tree
  - ◆ Uses Exploration Insights
- ➔ What Classifier to Use?
  - ◆ Random Forest Classifier
    - ● Decision Tree Ensemble



Random Forest Classifier

- Important Parameters
  - Number of Estimators (Decision Trees)
    - ↑ Estimators : ↑ Better Score : ↓ Efficiency
    - Set to 10
  - Maximum Features
    - ↓ Max : ↓ Variance : ↑ Bias : ↓ Overfitting
    - Set to 5 (Square Root # of Columns)
- Performance
  - 100% Testing Accuracy
  - Never-Before Seen Test Set
    - 2,301 Mushrooms
  - Simple Model
    - Exploration Phase Greatly Helped