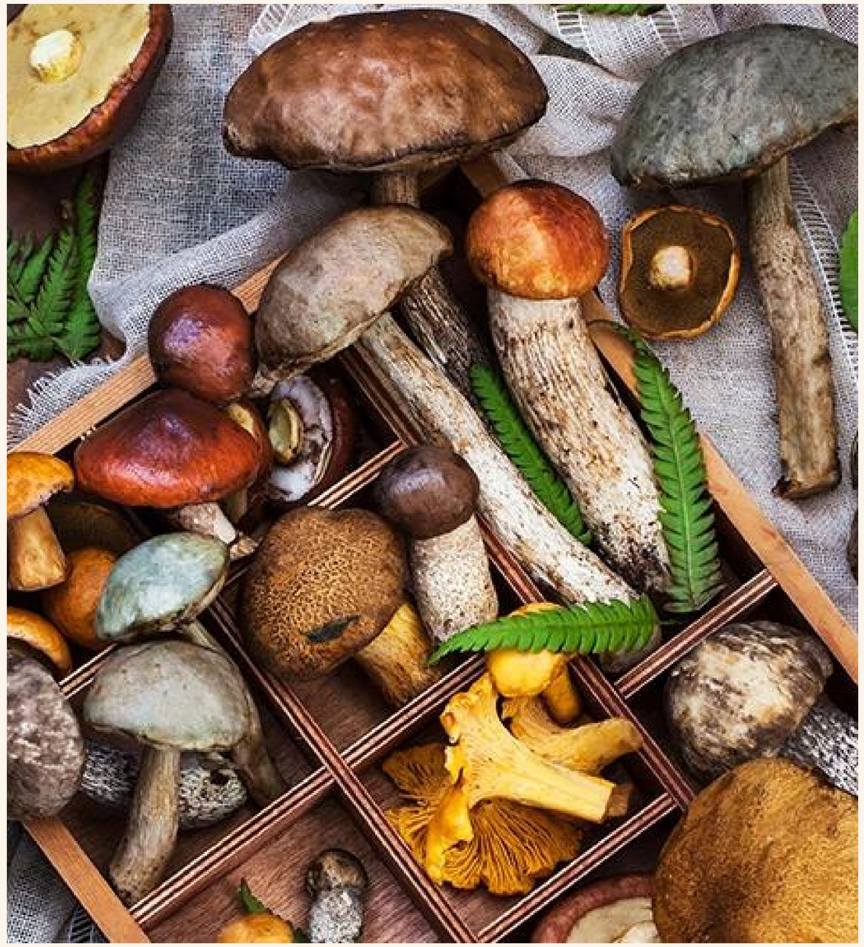
Mushroom Classification

Identifying whether mushrooms are edible or poisonous

IIM L EPDS Batch# 4 Group# 1

Bijoy Yohannan	Sr. Manager, Bosch
Gaurav Tripathi	Sr. Manager, Telstra
Manish Mehta	AVP, State Street
Mausam Kumar	Deputy Managar, BHEL
Pradeep Sarangdharan	Director, Pensalto Consulting
Rajesh Vasudevan	Director, Fiserv
Sakshi Mathur	Assoicate, Dutsche Bank
Sushant Parashar	Officer, Canara Bank



Agenda

- Introduction
- Problem Scope
- Data Source
- Mushroom Dataset
- Dataset Attribute Info
- ETL Journey & Learning
- Exporatory Data Analysis
- Pre-Processing
- Feature Selection
- Building Prediction Models
- Conclusion & Final Thoughts
- Q&A

Introduction

Ingestion of wild and potentially toxic mushrooms is common across the globe. Today most mushrooms are farmed commercially, but foraging for mushrooms is on the rise as a recreational endeavor. Mushroom poisonings can occur because of forager misidentification of a poisonous species as edible.

The goal of this project is to construct a model with high accuracy, precision, and recall, which can predict whether a mushroom is edible or poisonous.

Scope of Problem

- Of the 5000 mushroom species identified worldwide, about 3% are poisonous
- Mushroom poisoning is a critical health problem in many countries
- Most mushroom poisonings reported are accidental oral ingestion of poisonous mushrooms misidentified for edible species
- Mushroom foraging is rising in popularity

Important Stats

- 7428/year ingest poisonous mushrooms with minor harm
- 39/year experience major harm
- 2.9/year die from ingestion

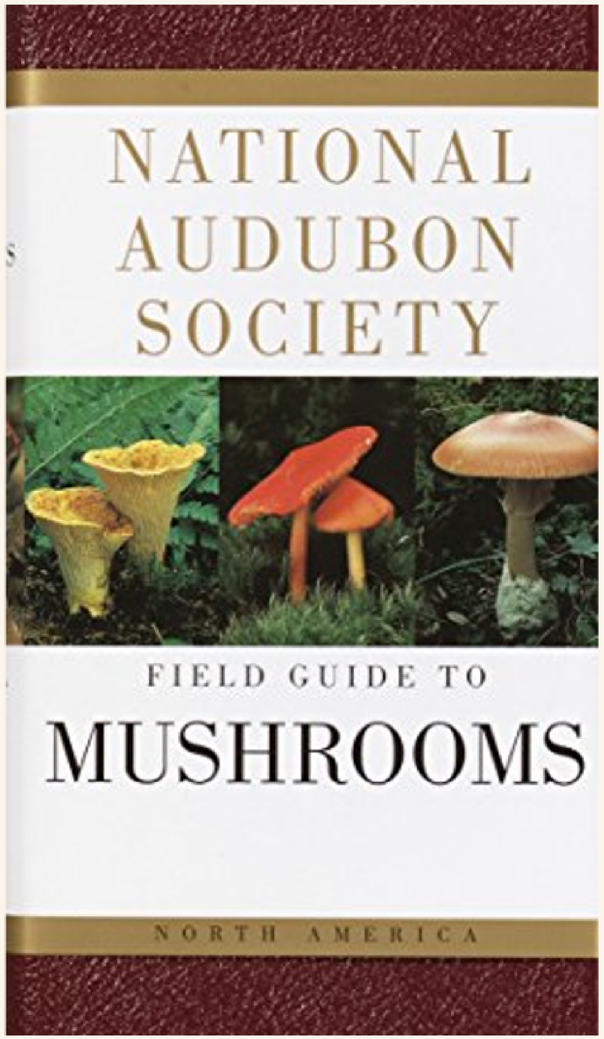
Will You eat this?

Deadly Galerina Mushroom

- Also known as funeral bell, deadly skullcap, autumn skullcap or deadly galerina, is a species of extremely poisonous mushroom

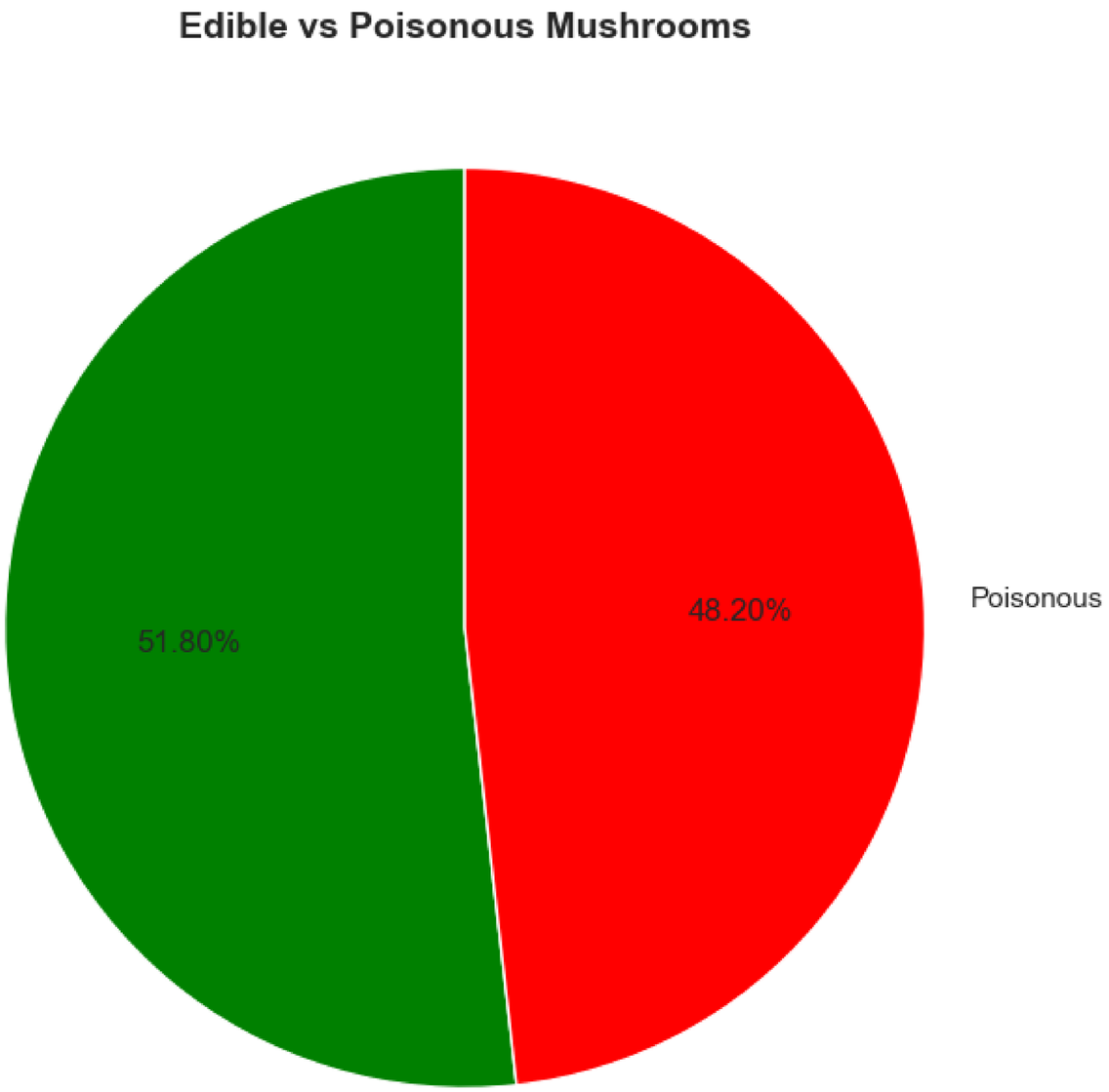
- Ingestion in toxic amounts causes severe liver damage with vomiting, diarrhea, hypothermia, and eventual death if not treated rapidly.

Data Source



- UCI Machine Learning Repository
- Data retrieved from The Audubon Society Field Guide to North American Mushrooms, published in 1981

Mushroom Dataset



Overview

- 22 Attributes: (Explained in detail in next slide)
- Class Attribute: Edible(e) or Poisonous(p)

Dataset Characteristics	Multivariate
Attribute Characteristics	Categorical
Number of Intances	8124
Number of Attributes	22

Dataset Attribute Information

22 Attributes

18 Intrinsic

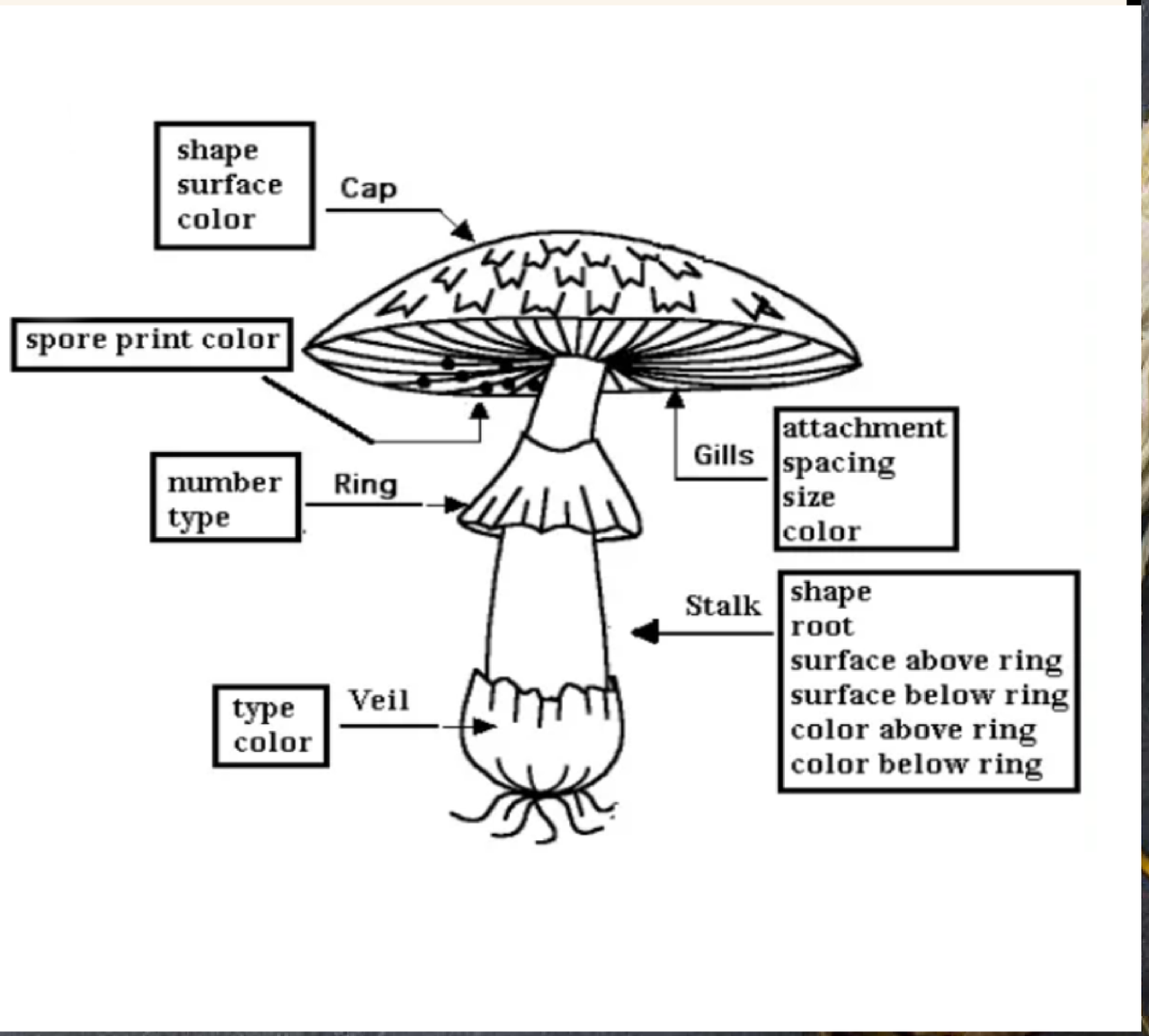
4 Others

1 Habitat

1 Population

1 Bruises

1 Odor



Dataset Attribute Information

	Column Name	Attribute Information
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

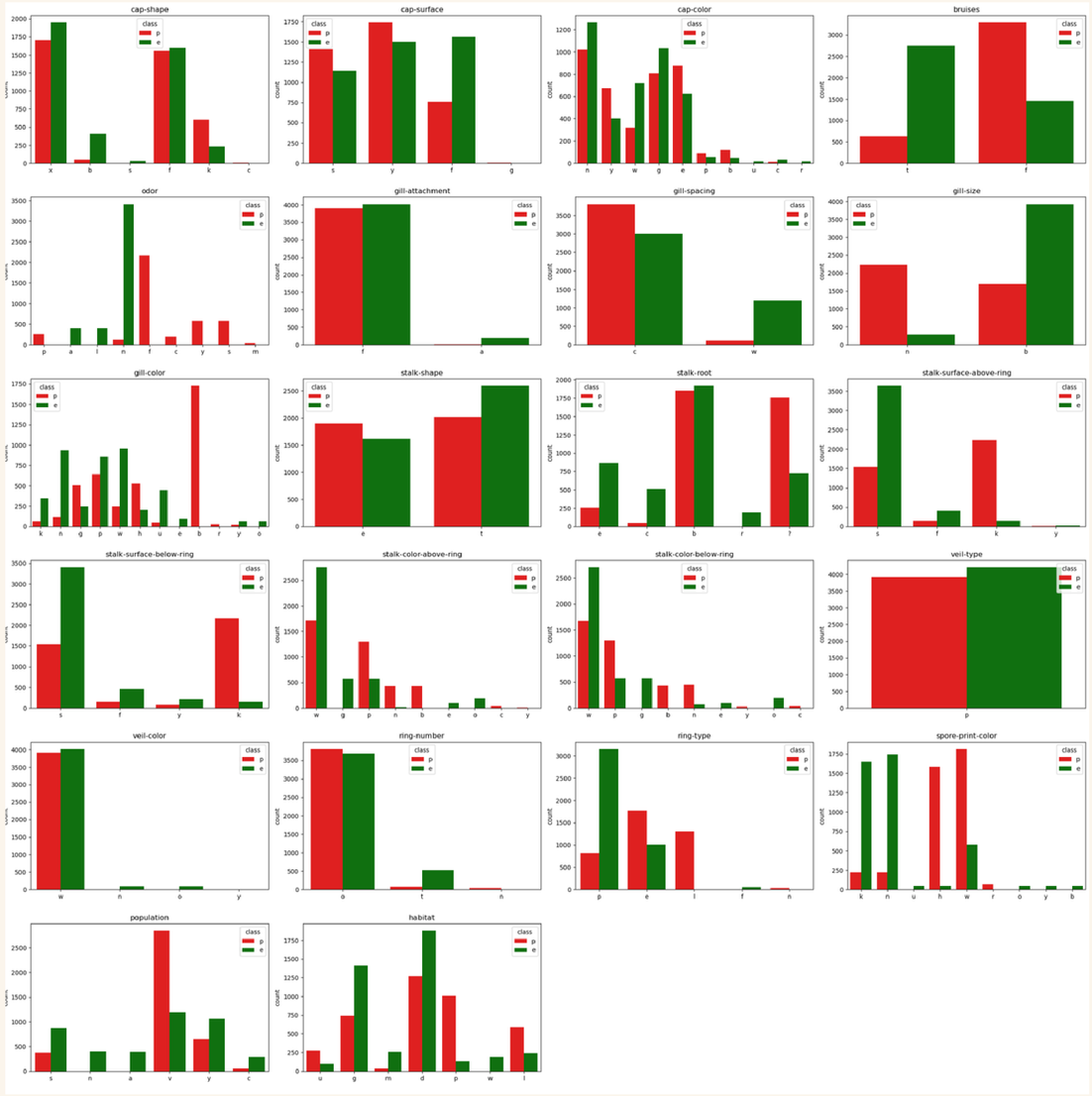
Will You eat this?

Morchella/True Morel/Guchhi

- It is one of the most readily recognized of all the edible mushrooms and highly sought after.

- Morel mushrooms contain a lot of vitamin D. They are also a low-fat, plant-based food that makes a great addition to a heart-healthy diet as an ingredient or as a meat substitute.

Exploratory Data Analysis

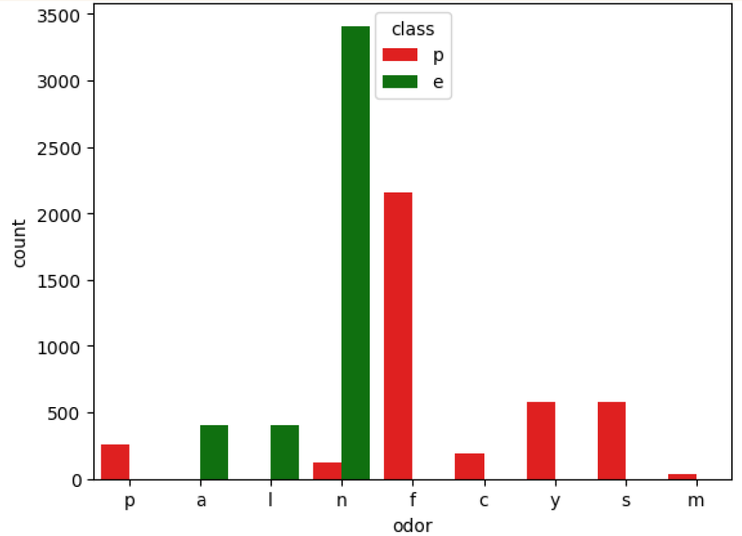


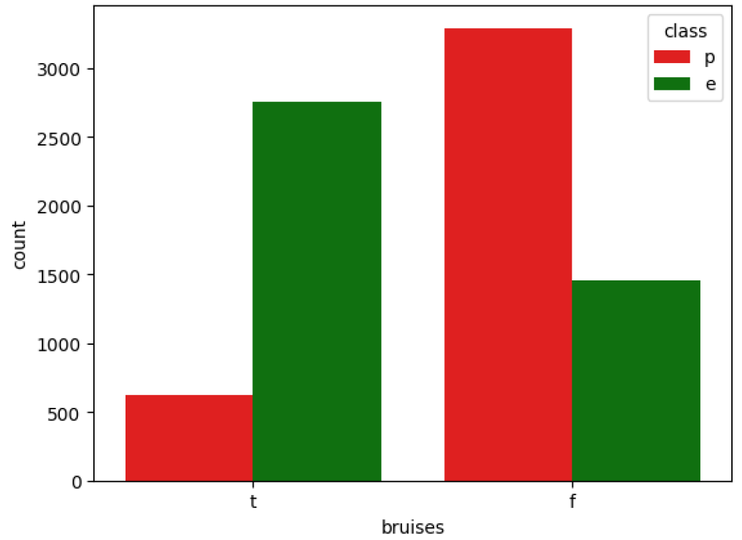
Observations

- **Stalk-root :**
  - On Close examination we discover that the "stalk-root" has "?" listed as one the the unique values.
  - There are 2,480 such records which is a substantial proportion of our data (30.53% of mushrooms with a missing stalk-root value).
- **Veil-type:**
  - Every veil-type is "partial" (as opposed to universal).
- Similarly, almost all **veil-color** are white.
- Almost all **gill-attachment** are "free".
- The vast majority of mushrooms in this sample only have "one" **ring-number.**

As such, these variables will likely not be as useful for our analysis.

Exploratory Data Analysis





- Only the poisonous mushrooms have unpleasant **odor** (e.g. foul, musty, spicy, etc.)
- Only the poisonous mushrooms have a "buff" **gill-color**; however, other colors seems to be less telling
- Only the poisonous mushrooms have a "large" **ring-type**
- The majority of black (k) and brown (n) **spore-print-color** are from edible mushrooms. The majority of chocolate (h) and white (w) spore print colors are from poisonous mushrooms
- The majority of poisonous mushrooms have a **population** of "several" (v). Only edible mushrooms have numerous (n) or abundant (a) populations
- If there are **bruises** on mushroom, it is highly likely to be an edible mushroom

Will You eat this?

Gyromitra esculenta/False Morel

- The false morel was the most common cause of mushroom-related poisoning in Poland after WWII.
- The name false morel is given to several species of mushroom which bear a resemblance to the highly regarded true morels of the genus Morchella.
- It is poisonous, even fatal, unless properly dried and boiled.

Pre-Processsing

Step 1: Feature Engineering & Label Encoding

- Recategorize odor under two broad umbrella – 'bad smell' and 'not bad smell'

odor	almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s

- Convert all variables to numerical (necessary for machine learning algorithms)

```
In [15]: from sklearn.preprocessing import LabelEncoder
         # Instantiate the label encoder
         le = LabelEncoder()

         # Create a copy
         mushroom_num = mushroom.copy()

         # Iterate through columns
         for col in mushroom_num.columns:
             if mushroom_num[col].dtypes == 'object':
                 # Complete transformation
                 mushroom_num[col] = le.fit_transform(mushroom_num[col])
```

```
In [16]: mushroom['class'].value_counts()

Out[16]: e    4208
         p    3916
         Name: class, dtype: int64
```

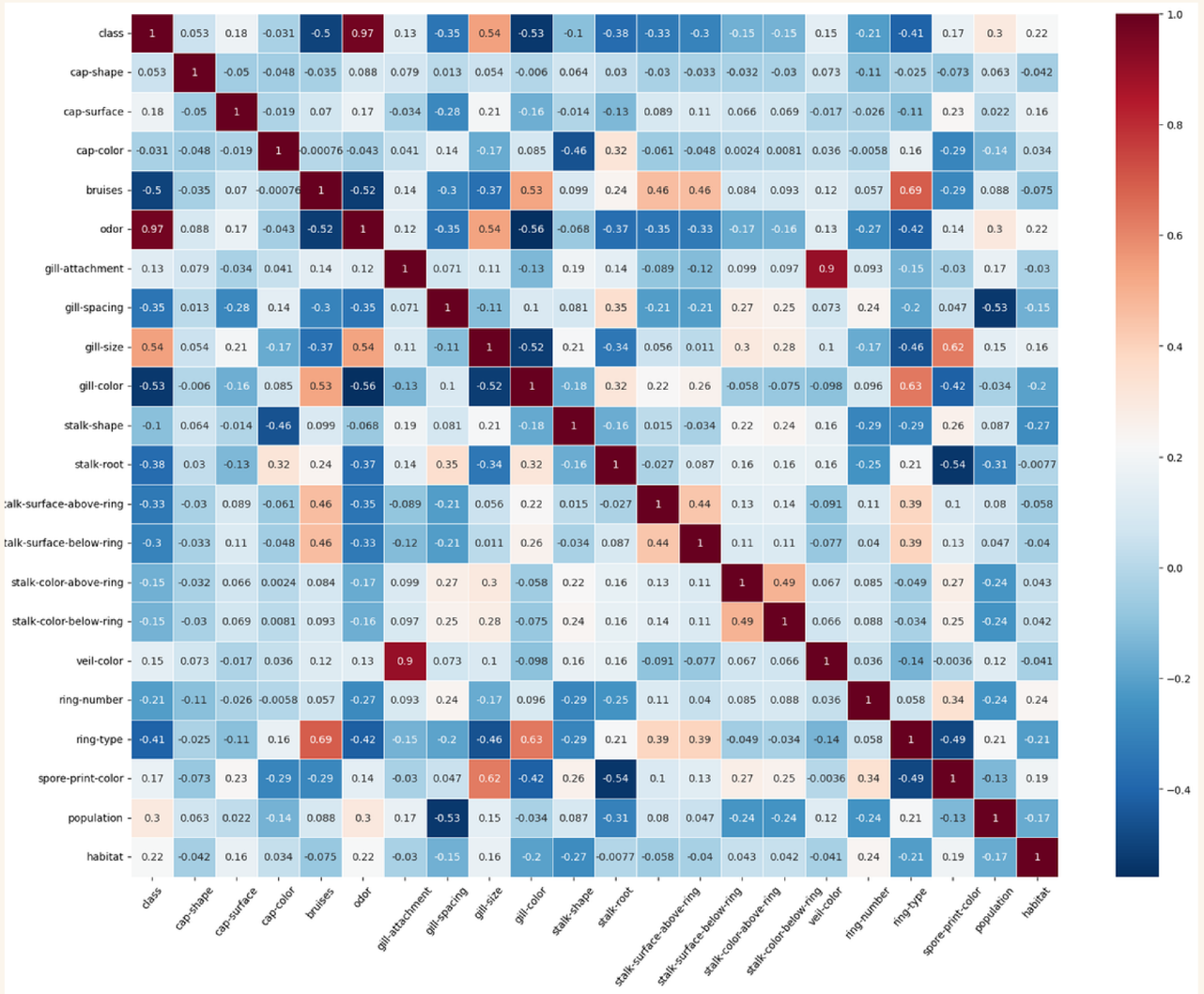
```
In [17]: mushroom_num['class'].value_counts()

Out[17]: 0    4208
         1    3916
         Name: class, dtype: int64
```

Pre-Processsing

Step 2: Feature Importance

- As all the features are categorical in nature, we ran a chi square test with alpha = 0.05
  - veil-type is NOT important. (p = 1.0, chi2_stat=0.0)
- After dropping veil-type, still we have 21 categorical features to sort from. So next we checked for Correlation Coefficient.

	Feature	Corr_coeff	VIF
0	odor	0.970814	7.709720
1	gill-size	0.540024	6.568617
2	gill-color	0.530566	6.291863
3	bruises	0.501530	9.655519
4	ring-type	0.411771	13.915805
5	stalk-root	0.379361	7.676081

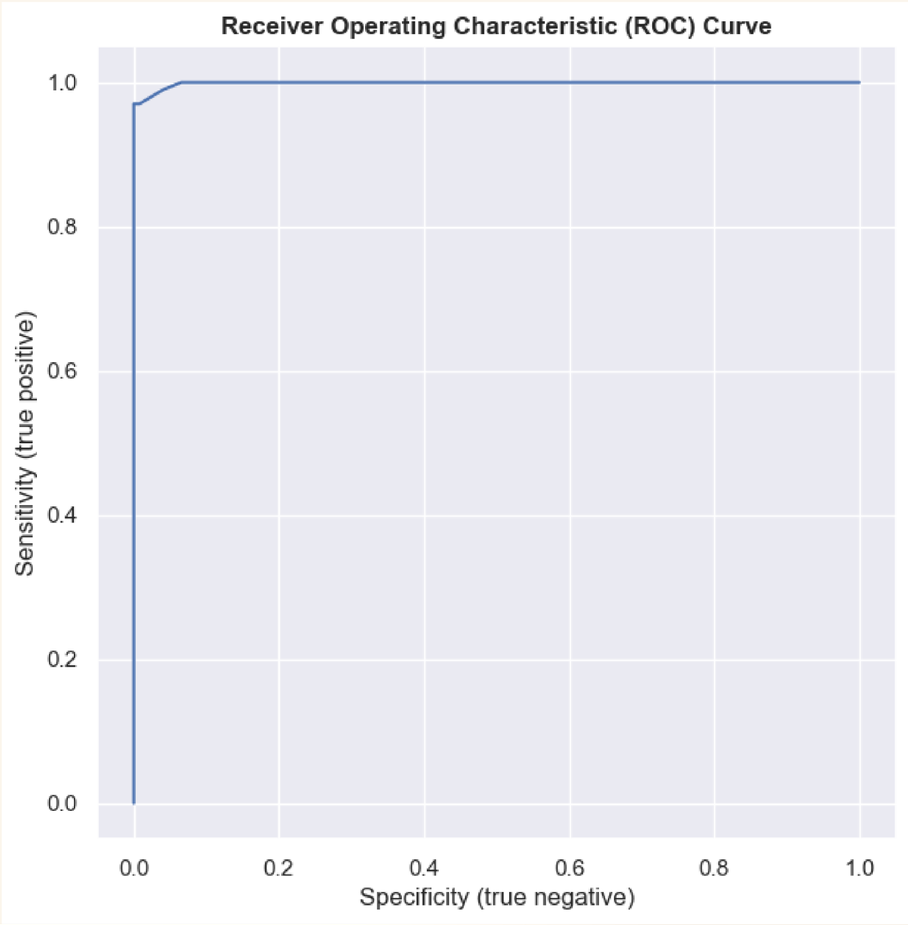


Pre-Processsing

Step 3: Feature Selection and Standardization

Overfitting

- Selected set of 10 features, which cover all intrinsic features as well as extrinsic one.
- We were easily achieving accuracy of 100%.
- We ran iterations of Logistic Regression model and in each run we kept reducing the features.
- Final Features selected:
  - gill-color
  - odor
  - bruises
  - habitat



Will You eat this?

Oyster Mushroom/ Hiratake/Pearl Mushroom

- The oyster mushroom is one of the more commonly sought wild mushrooms
- Oyster mushrooms are a superfood because they have nutrients, vitamins, minerals, antioxidants and bioactive compounds.
- They contain beta-glucans, which is best for protecting you against short term and long term illness by boosting your immunity.

Data Model 1 - Logistic Regression

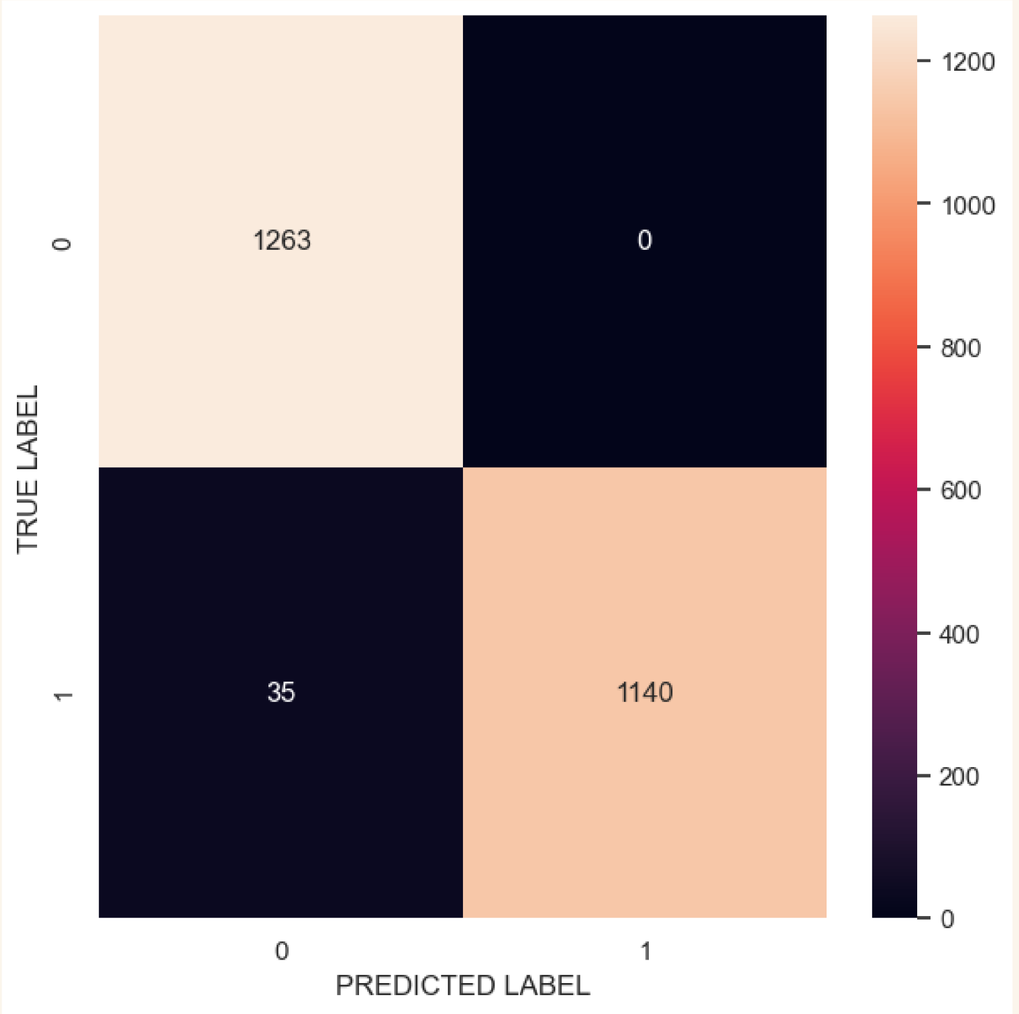
Model Evaluation Scores :

- Logistic Regression training accuracy: 98.51%
- Computation time taken for Logistic Regression: 0.02 seconds
- Area Under the Curve (AUC in %) : 98.5%

	precision	recall	f1-score	support
0	0.97	1.00	0.99	1263
1	1.00	0.97	0.98	1175
accuracy			0.99	2438
macro avg	0.99	0.99	0.99	2438
weighted avg	0.99	0.99	0.99	2438

Obervations :

- High precision model
- Low cost model
- Need to improve recall



Will You eat this?

Destroying Angel Mushroom

- The name destroying angel applies to several similar, closely related species of deadly all-white mushrooms in the genus Amanita.

- The destroying angel and the death cap account for the overwhelming majority of deaths due to mushroom poisoning.

- Symptoms do not appear for 5 to 24 hours, by which time the toxins may already be absorbed and the damage is irreversible.

Data Model 2- KNN Classification

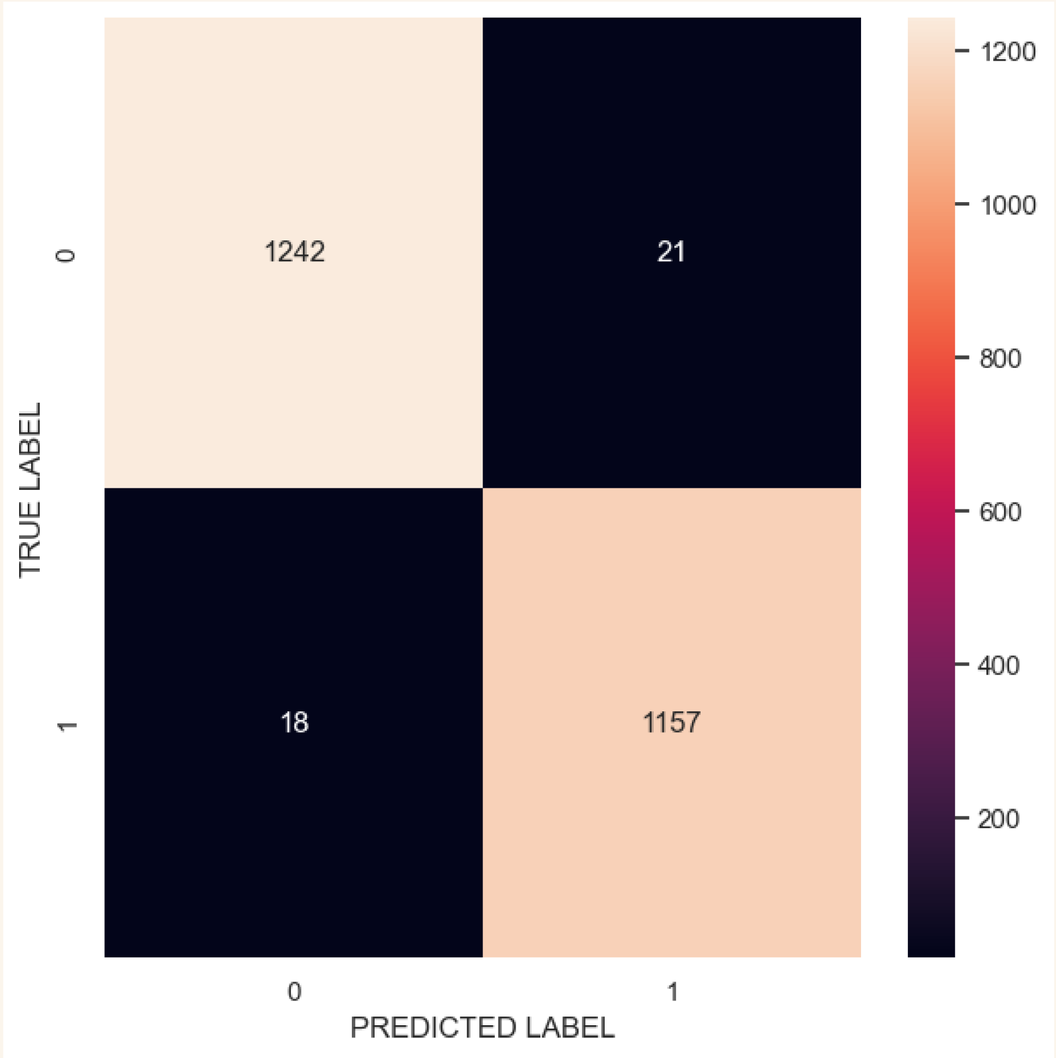
Model Evaluation Scores :

- KNN Model training accuracy: 99.19%
- Computation time taken for KNN Classificaton: 0.24 seconds
- Area Under the Curve (AUC in %) : 98.4%

	precision	recall	f1-score	support
0	0.99	0.98	0.98	1263
1	0.98	0.98	0.98	1175
accuracy			0.98	2438
macro avg	0.98	0.98	0.98	2438
weighted avg	0.98	0.98	0.98	2438

Obervations :

- Higher precision & recall then Logistic Regression
- Low Cost Model
- While recall improved precision reduced



Will you eat this?

Shiitake Mushroom

- It is considered a medicinal mushroom in some forms of traditional medicine.

- Known to reduce cholesterol levels in the blood. They also contain beta-glucans that reduce inflammation and help prevent the intestines from absorbing cholesterol. Enhnances immunity.

Data Model 3 - Decision Tree Classification

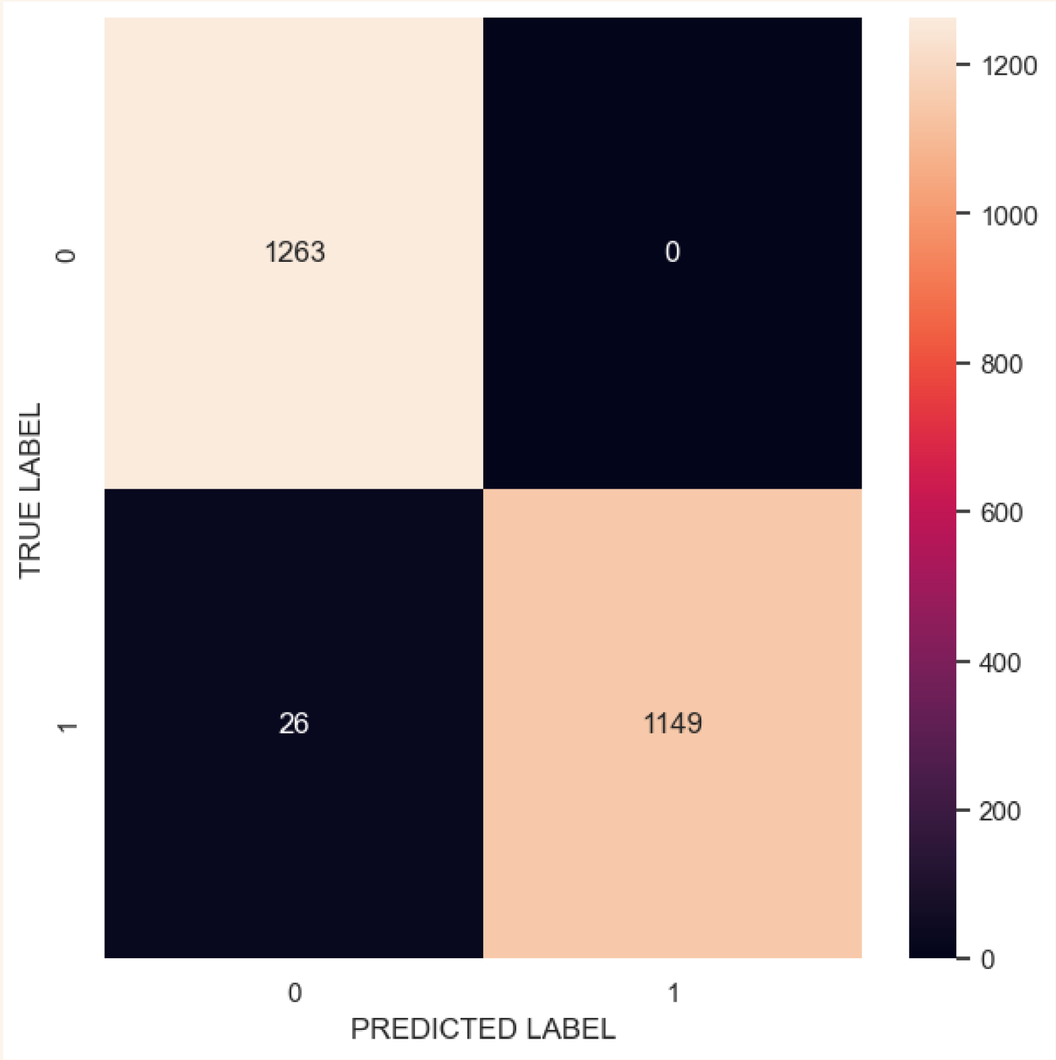
Model Evaluation Scores :

- Decision Tree training accuracy: 99.16%
- Computation time taken for Decision Tree: 0.34 seconds
- Area Under the Curve (AUC in %) : 98.9%

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1263
1	1.00	0.98	0.99	1175
accuracy			0.99	2438
macro avg	0.99	0.99	0.99	2438
weighted avg	0.99	0.99	0.99	2438

Obervations :

- High precision model
- Low Cost
- Need to increase recall



Will you eat this?

Enoki Mushroom

- Enoki mushrooms are considered as among the costliest mushrooms which can be grown at home using growing kits.

- They also have anticancer, anti-allergy, antibacterial, antiviral, and anti-inflammatory properties, which protect and boost your immunity

Data Model 4 - Random Forest Classification

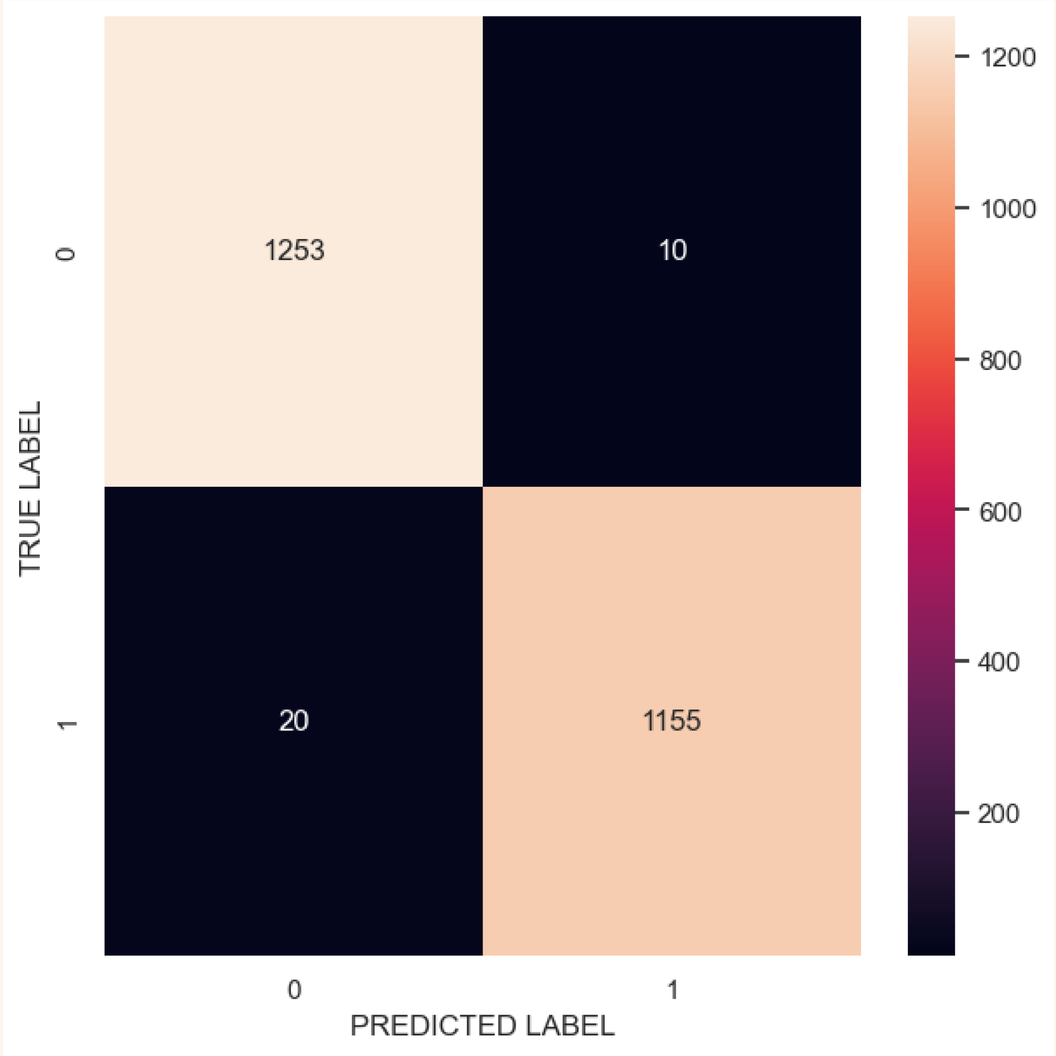
Model Evaluation Scores :

- Random Forest training accuracy: 99.2%
- Computation time taken for Random Forest: 8 Seconds
- Area Under the Curve (AUC in %) : 98.7%

	precision	recall	f1-score	support
0	0.98	0.99	0.99	1263
1	0.99	0.98	0.99	1175
accuracy			0.99	2438
macro avg	0.99	0.99	0.99	2438
weighted avg	0.99	0.99	0.99	2438

Observations :

- A good trade off between precision and recall
- Higher Cost
- Ensemble method helps against overfitting

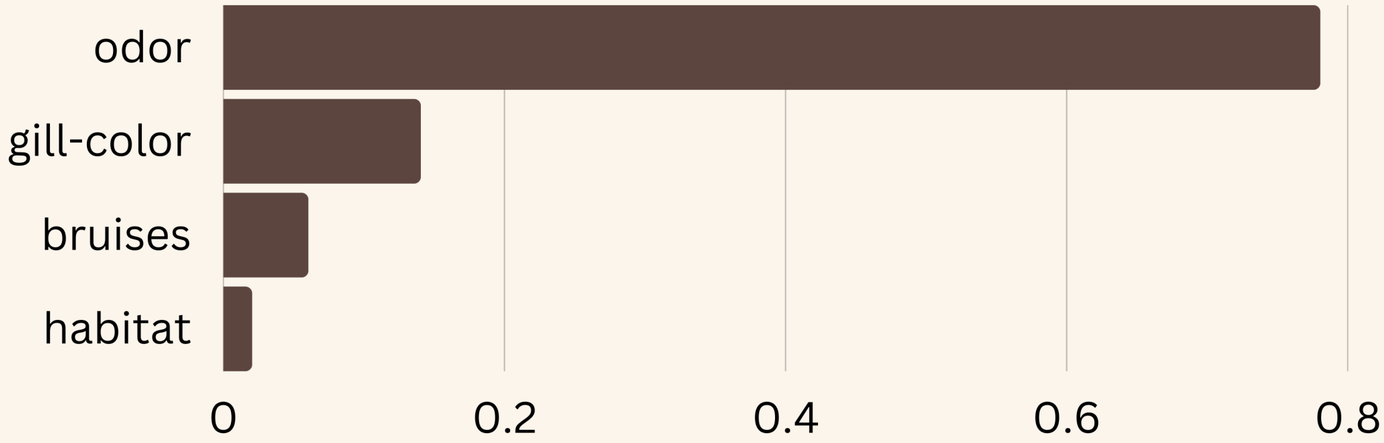


Best Model

Random Forest

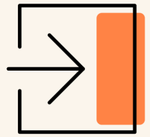
- Best overall accuracy score
- Best AUC score
- While tied with KNN on scoring, ensemble methods are more robust against over-fitting

How to identify the edible one then ?



Analysis and Conclusion

Exploratory analysis can help you discover predictive variables



It can be beneficial to regroup categorical variables into smaller categories to avoid overfitting.

Sometimes, different models will have very similar scoring, if the individual variables are highly predictive

Thank you!

Do you have any questions?