Mushroom Classification

Ahmed Mohammed Gaber

Important Question

How can we prevent acidenets of poisoning while shrooming ?

At least three people have died and hundreds were sickened in France in the past two months after eating wild mushrooms.

Poisoning risks include confusing edible types with toxic species

This project will help in classifying the Poisionous Mushroom from the edible ones

Index

1. Exploring Data
2. EDA
3. Data Preprocessing
4. Training the models
5. Evaluating the modes
6. Choosing the best model
7. Feature selection
8. Deploying the model

1 Exploring Data

Eploration results

- mushroom hunting is enjoying new peaks in popularity
- The dataset contains 8124 record
- There are 23 features descriping the size and color and other features of the mushroom
- The dataset does't contain Null values
- All the entries of type string

I noticed that in the satlk-root feature there is an unknown value ("?") that keeps showing

since a lot of records take this value it's better that we keep it (there exists some situations when we don't know the root of the plant)

```
print(df['stalk-root'].value_counts())
```

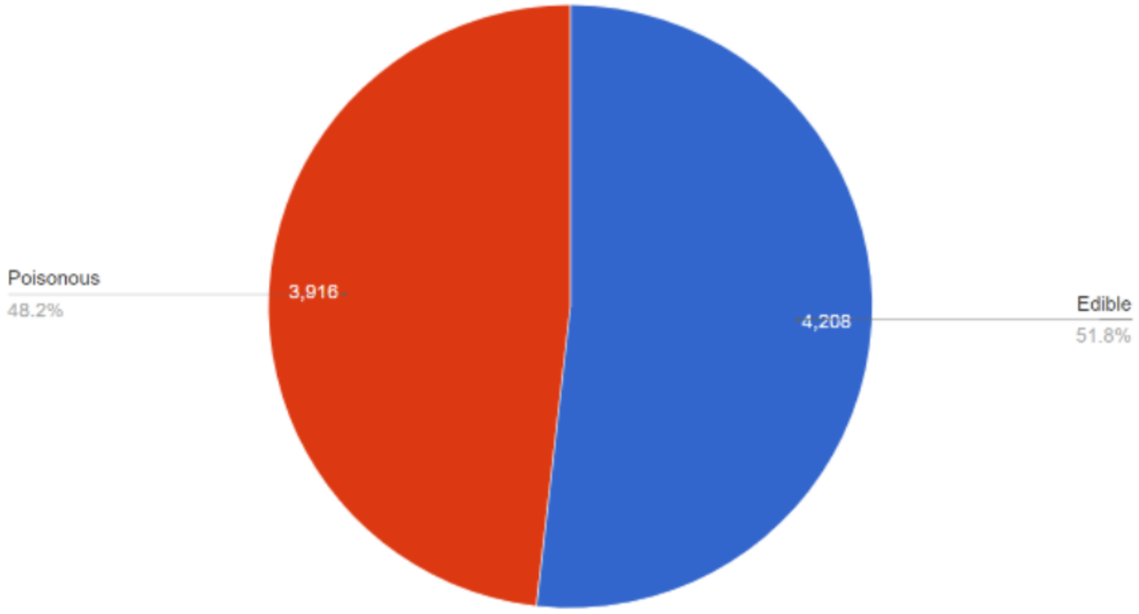
```
bulbous     3776
?           2480
equal       1120
club         556
rooted       192
Name: stalk-root, dtype: int64
```

It seems that the column "veil-type" has only one value so dropping it won't affect our work

```
print(df['veil-type'].value_counts())
df.drop('veil-type',axis = 1, inplace=True)
```

```
partial    8124
Name: veil-type, dtype: int64
```

2 Exploratory Data Analysis

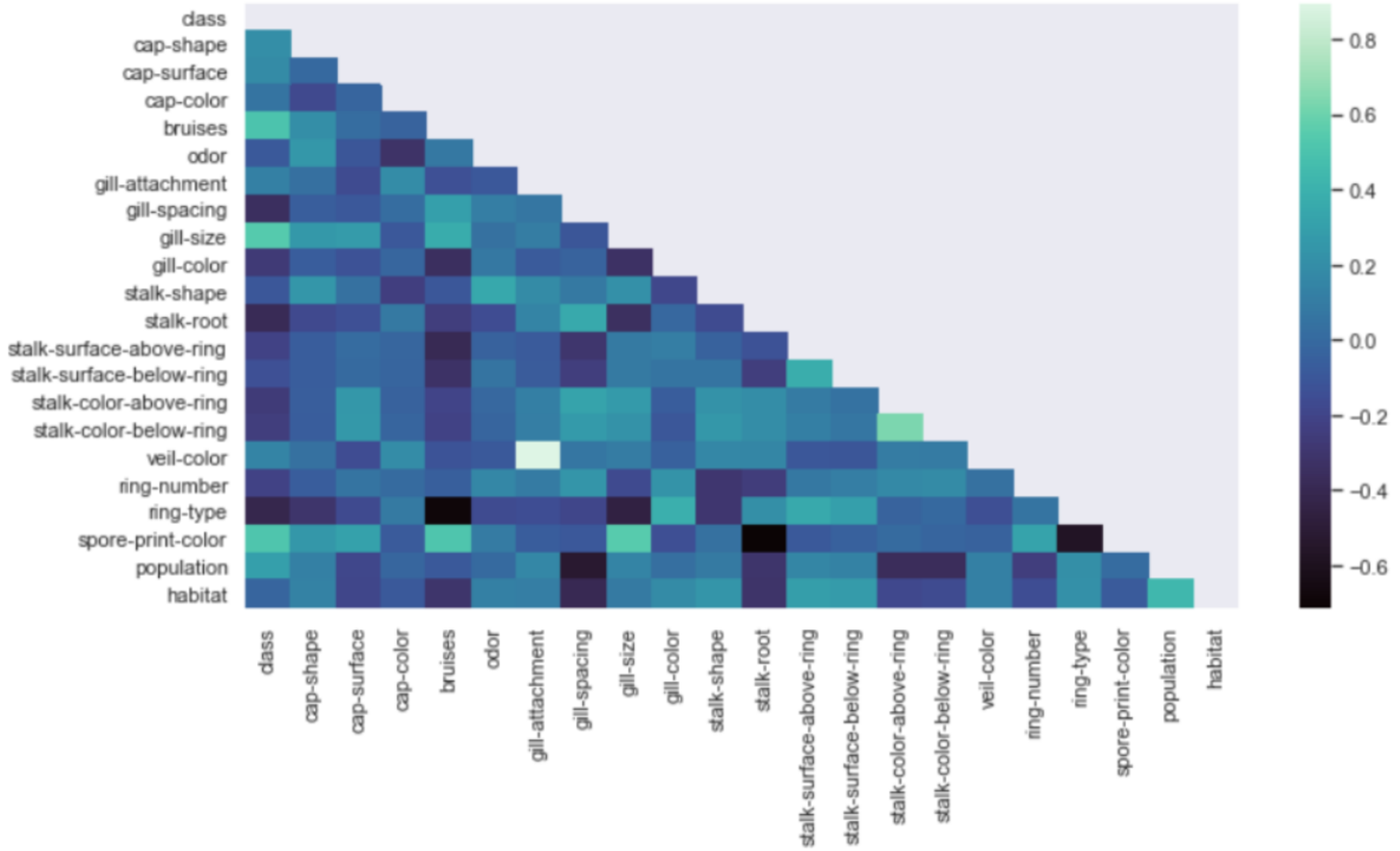


Checking for Imbalance

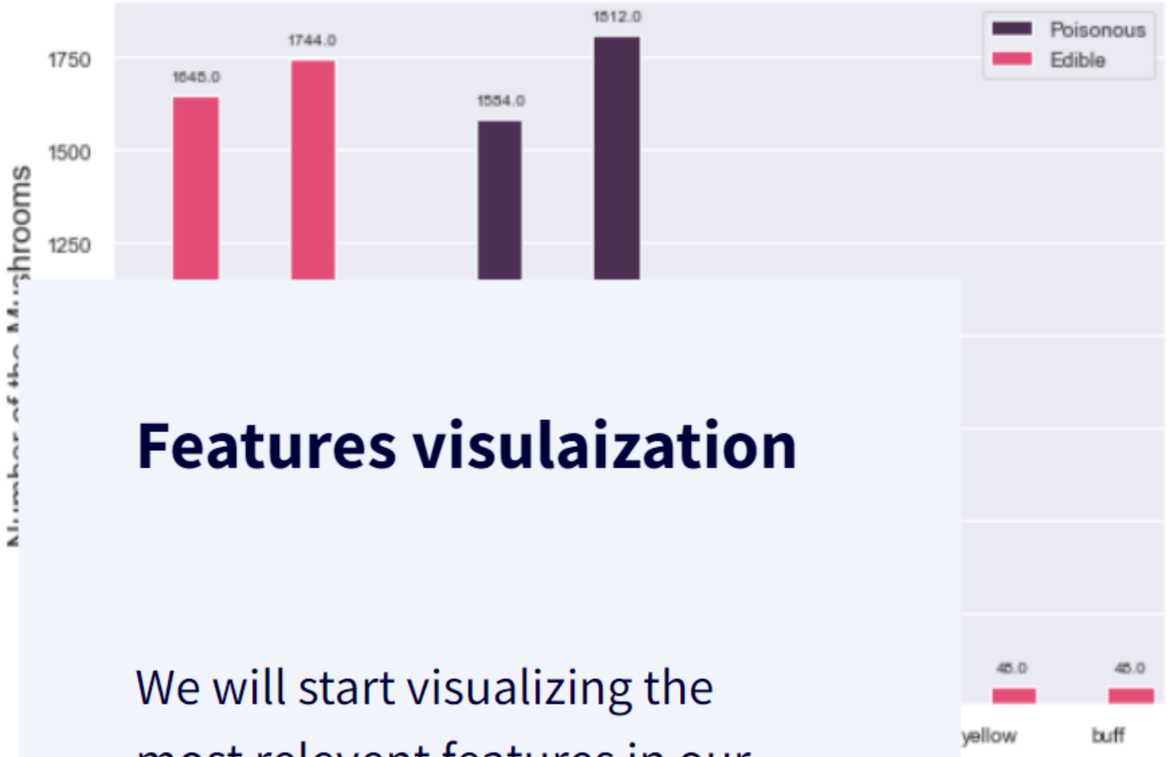
In order to get acurate results and acurate measures for acuurcy we have to check for imbalance in the dataset

We notice that don't need to use oversampling to handle imbalance , the data is well balanced

Correlation between features

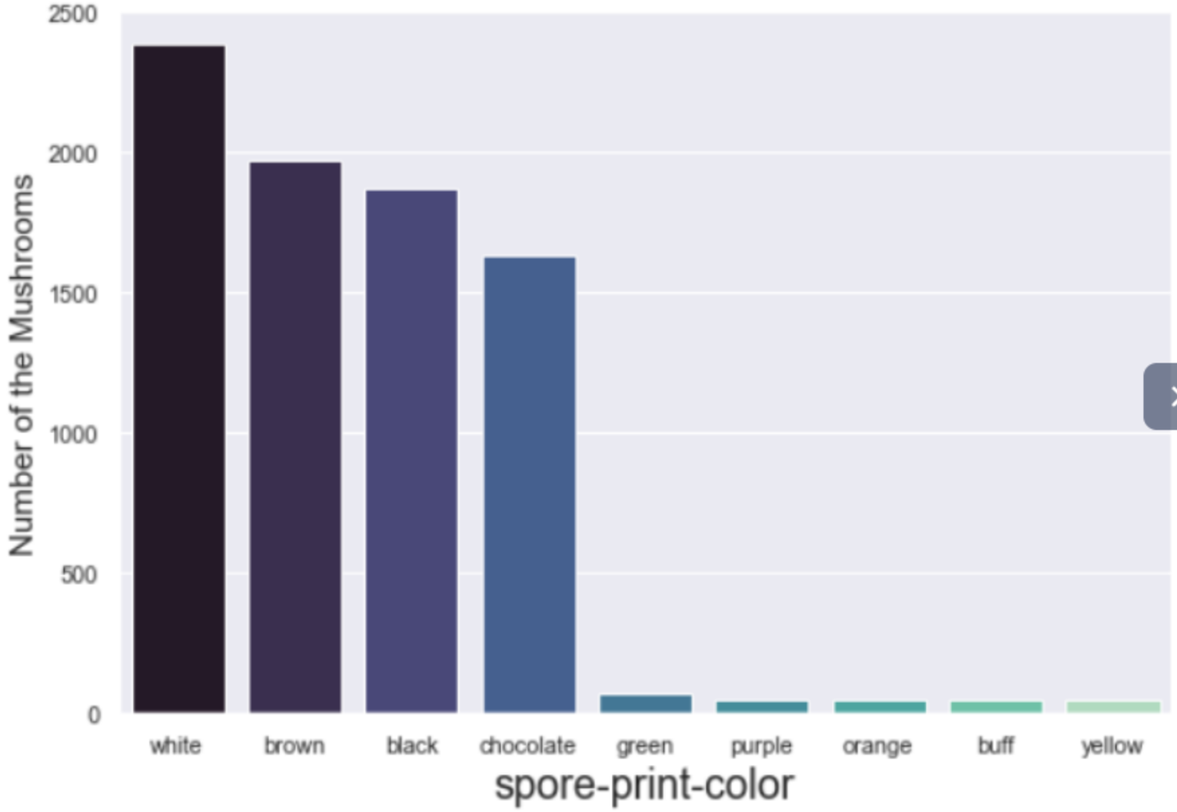


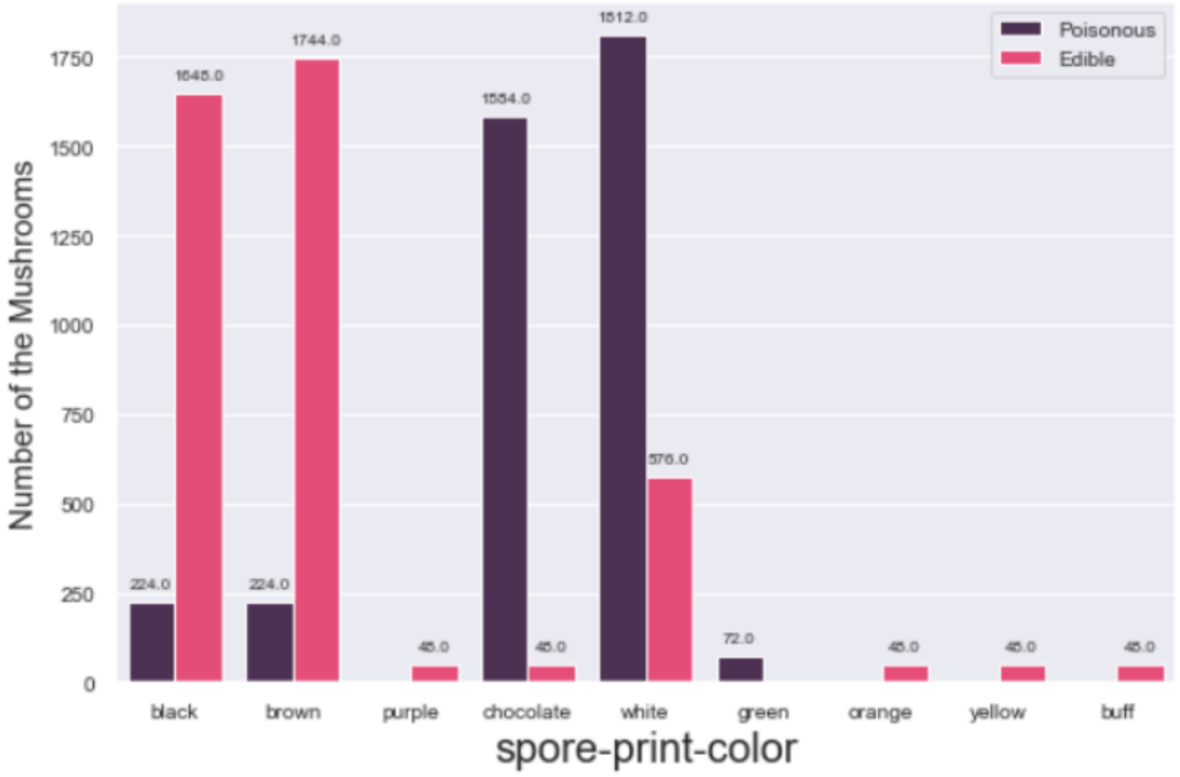
From the correlation heat map we can see that there is no dominanting feature and pretty much all the features correlate to the debendent variable "class"

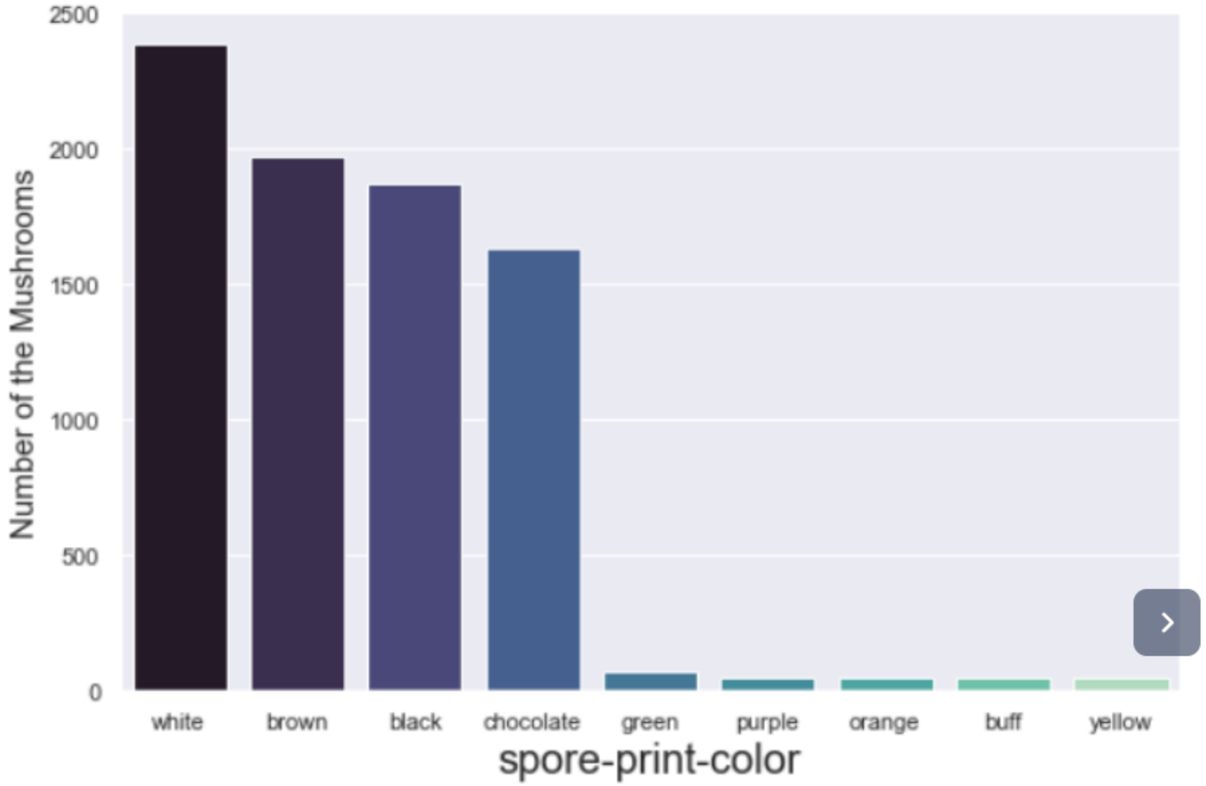


Features visulaization

We will start visualizing the most relevent features in our dataset to get more insights



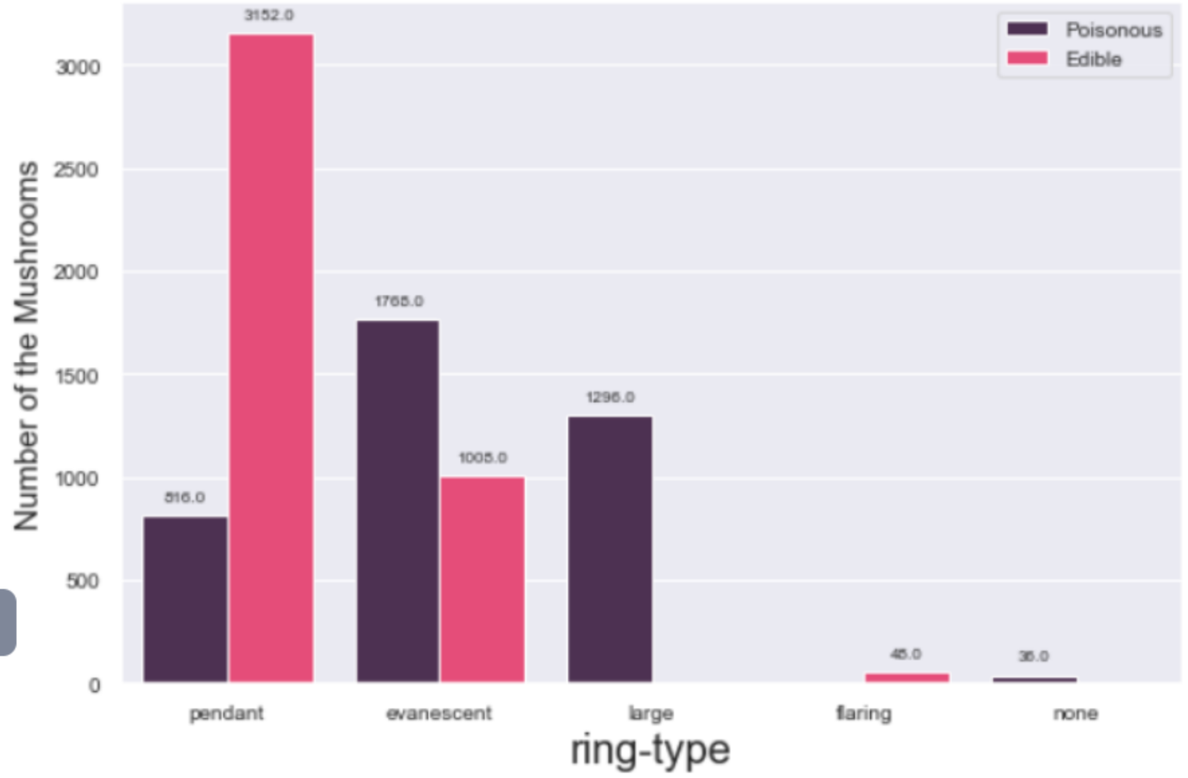


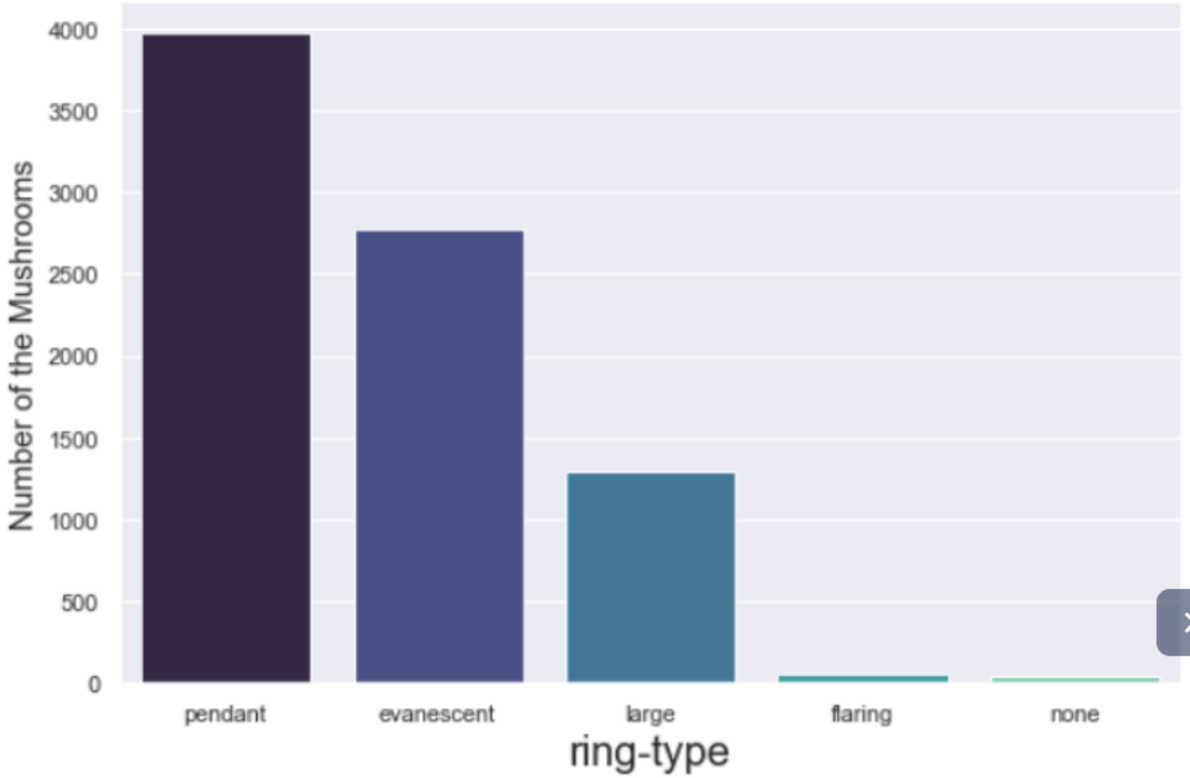


Spore print color with class Vs count

We notice that the colors Black and brown are more likely to be edible and they are the second and the third most frequent colors

On the other side we see that the colors Chocolate and White are more liekely to be poisionous

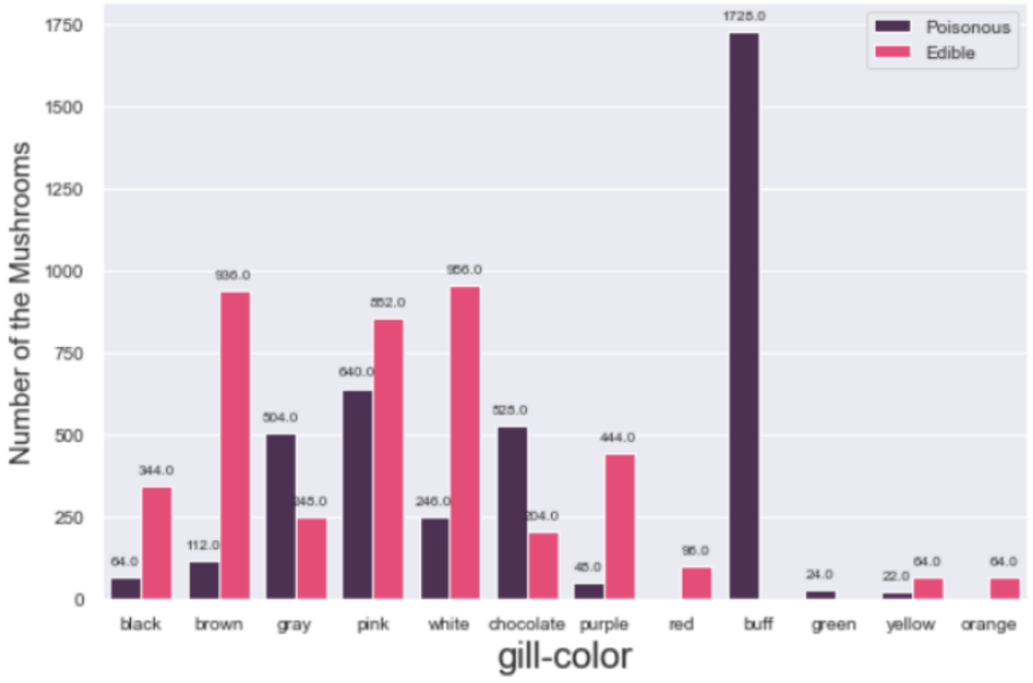


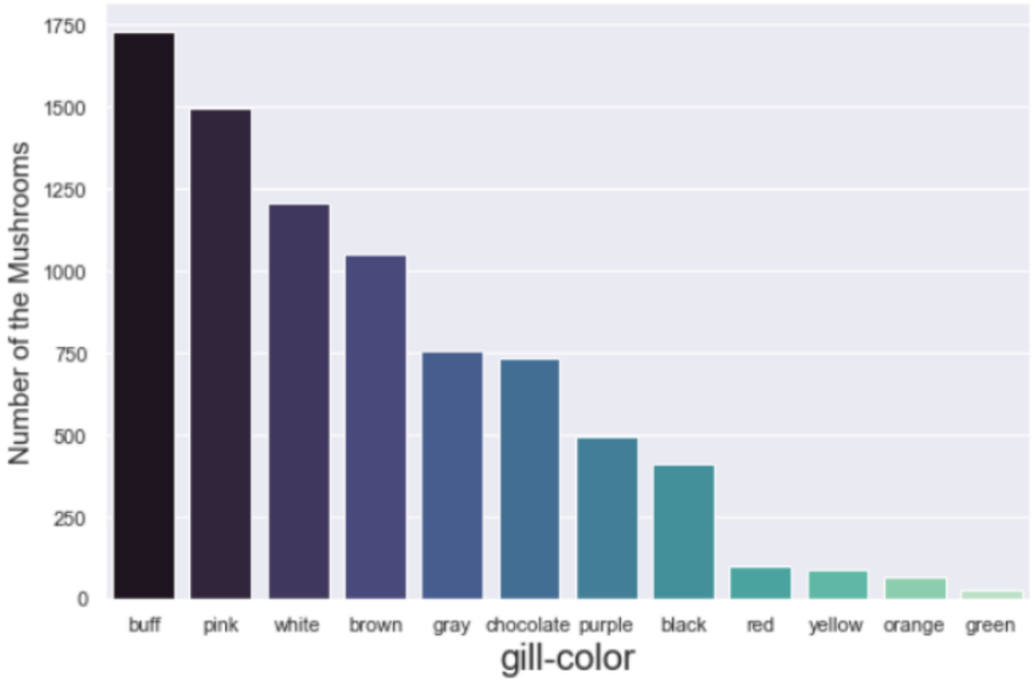


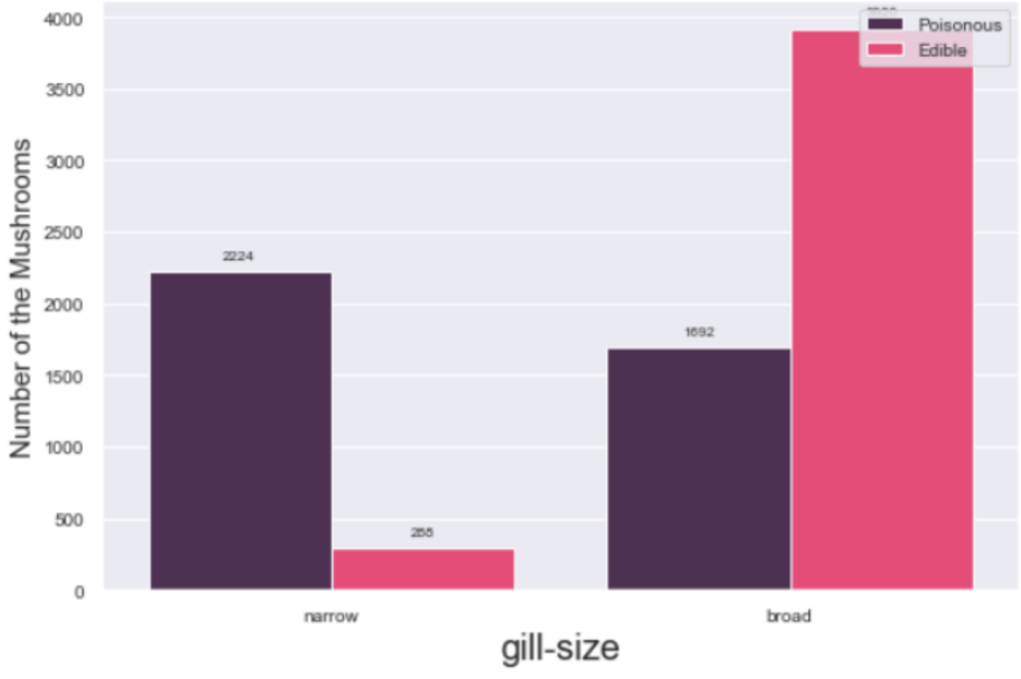
Ring type with class Vs count

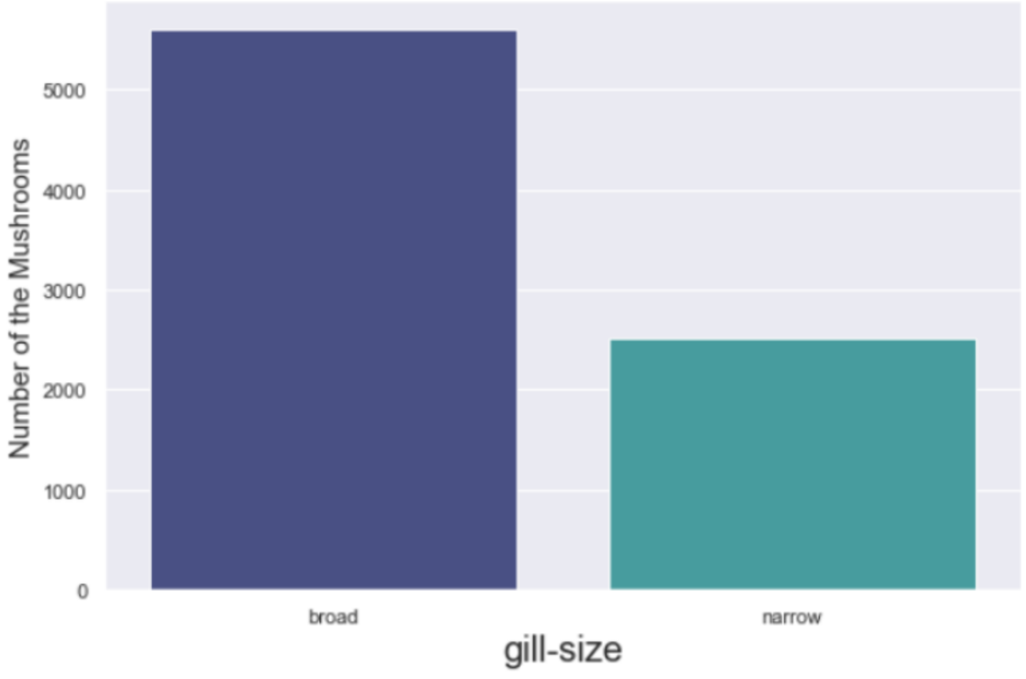
Here the "pendant" ring type is the most frequent type and more likely to be edible

let the chart speak for the rest


Poisonous
Edible
Number of the Mushrooms
gill-color
black
brown
gray
pink
white
chocolate
purple
red
buff
green
yellow
orange
64.0
344.0
112.0
936.0
504.0
345.0
640.0
552.0
246.0
956.0
525.0
204.0
45.0
444.0
96.0
1725.0
24.0
22.0
64.0
64.0


Number of the Mushrooms
gill-color
buff
pink
white
brown
gray
chocolate
purple
black
red
yellow
orange
green


Poisonous
Edible
Number of the Mushrooms
gill-size
narrow
broad
2224
255
1692


Number of the Mushrooms
gill-size
broad
narrow

3 Data Preproccesing

Encoding

In order to convert the categorical variables into numerical variables allowing us to do numeric operatins on them , I used Label Encoding in this process and I saved the encoded values in a dictionary so that I can refer to it in the later process

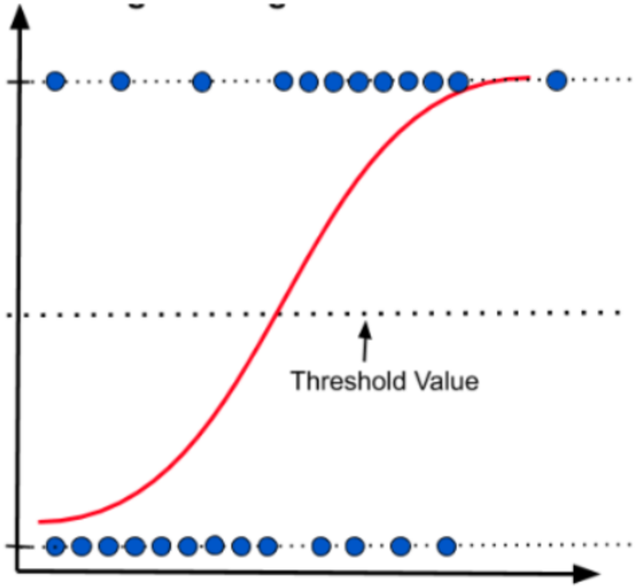
Feature Scaling

I scaled the features values in the dataset so all the features have a unit variance
This process eases the convergence of the machine learning algorithms

I used the Standard Scaling method to rescale the features

4 Training & Evaluating The Models

The trained models

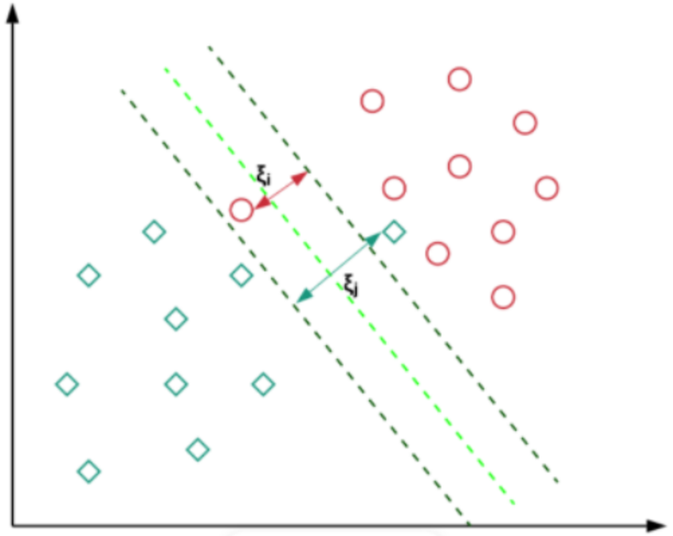


Logistic Regression

avg training time: 23.0 ms

avg testing time: 1.4 ms

avg teseting score: 94.9%

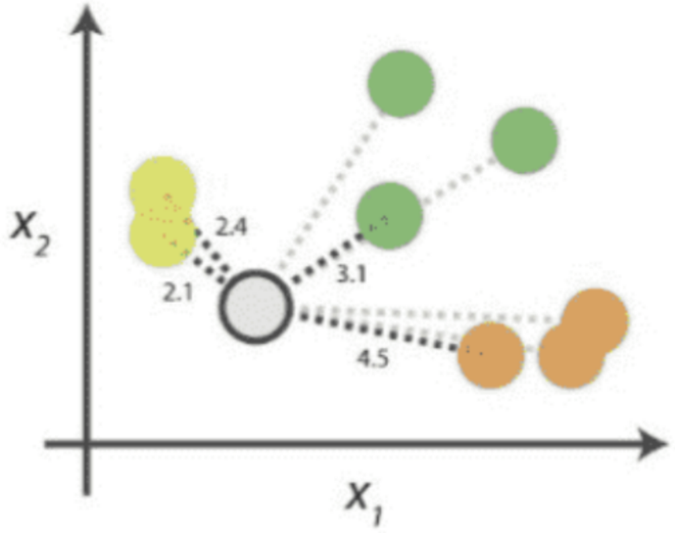


Support Vector Classifier

avg training time: 123.8 ms

avg testing time: 22.9 ms

avg teseting score: 100.0%



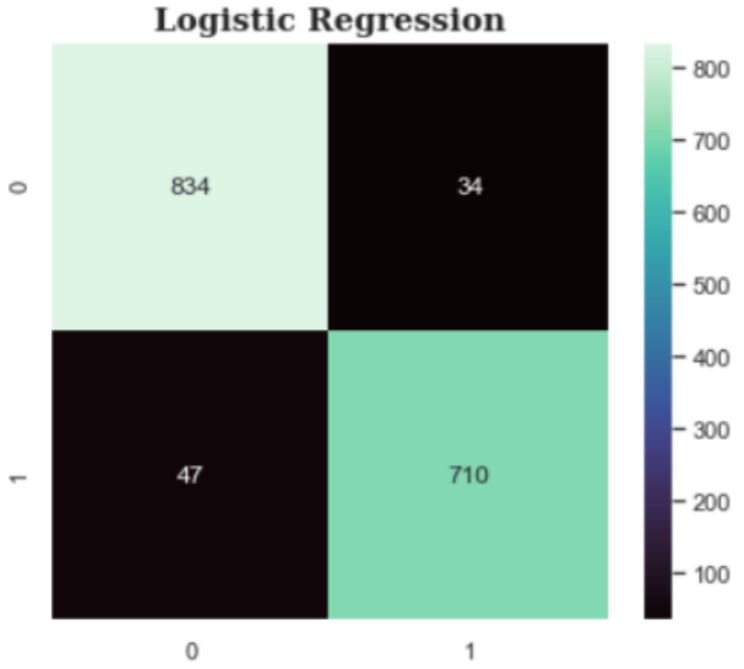
K-Nearest Neighbors

avg training time: 2.5 ms

avg testing time: 81.5 ms

avg teseting score: 99.98%

Confusion matrecies

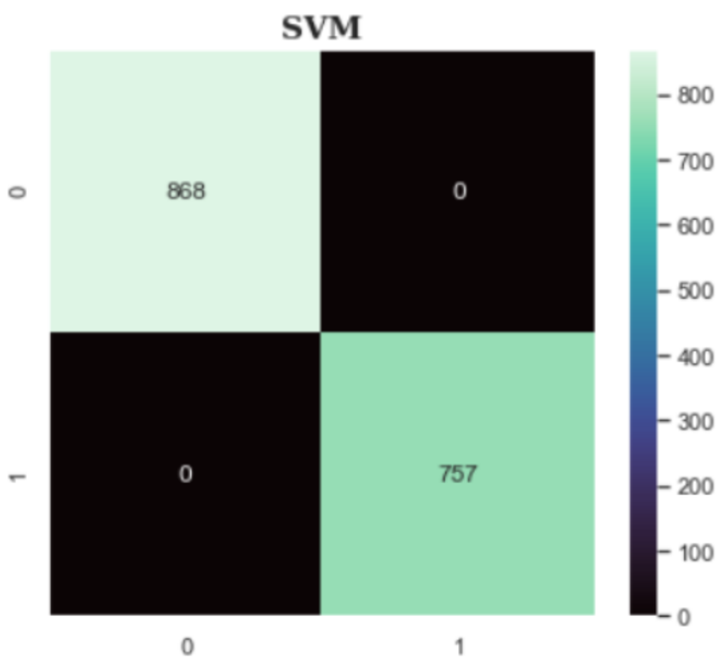


Logistic Regression

avg training time: 23.0 ms

avg testing time: 1.4 ms

avg teseting score: 94.9%

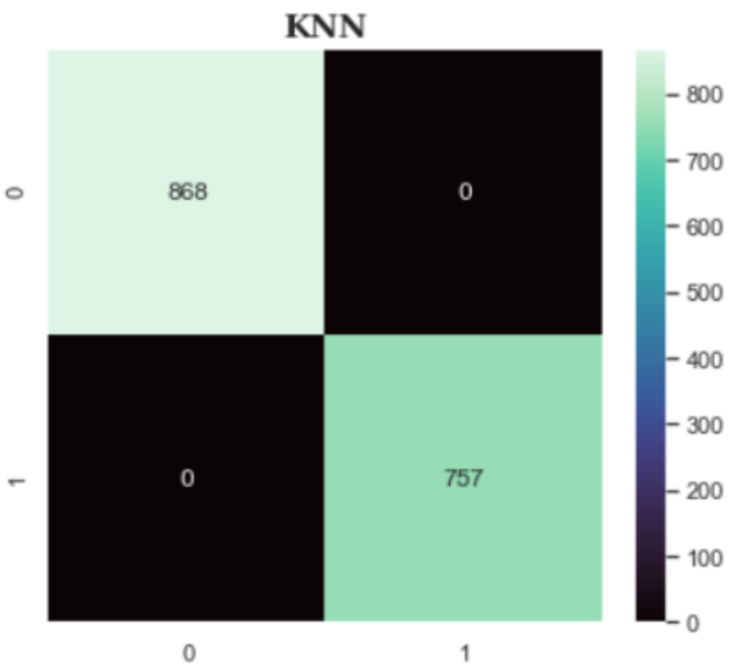


Support Vector Classifier

avg training time: 123.8 ms

avg testing time: 22.9 ms

avg teseting score: 100.0%



K-Nearest Neighbors

avg training time: 2.5 ms

avg testing time: 81.5 ms

avg teseting score: 99.98%

Classification report (SVM)

	precision	recall	f1-score	support
0	1.00	1.00	1.00	868
1	1.00	1.00	1.00	757
accuracy			1.00	1625
macro avg	1.00	1.00	1.00	1625
weighted avg	1.00	1.00	1.00	1625

Model Selection

- I Decided to choose the SVM model
- It's the most acurate model with 100% Precision and Recall
- There is no overfitting (the training and testing accurcy score both 100%)

5 Feature Selection & Deployment

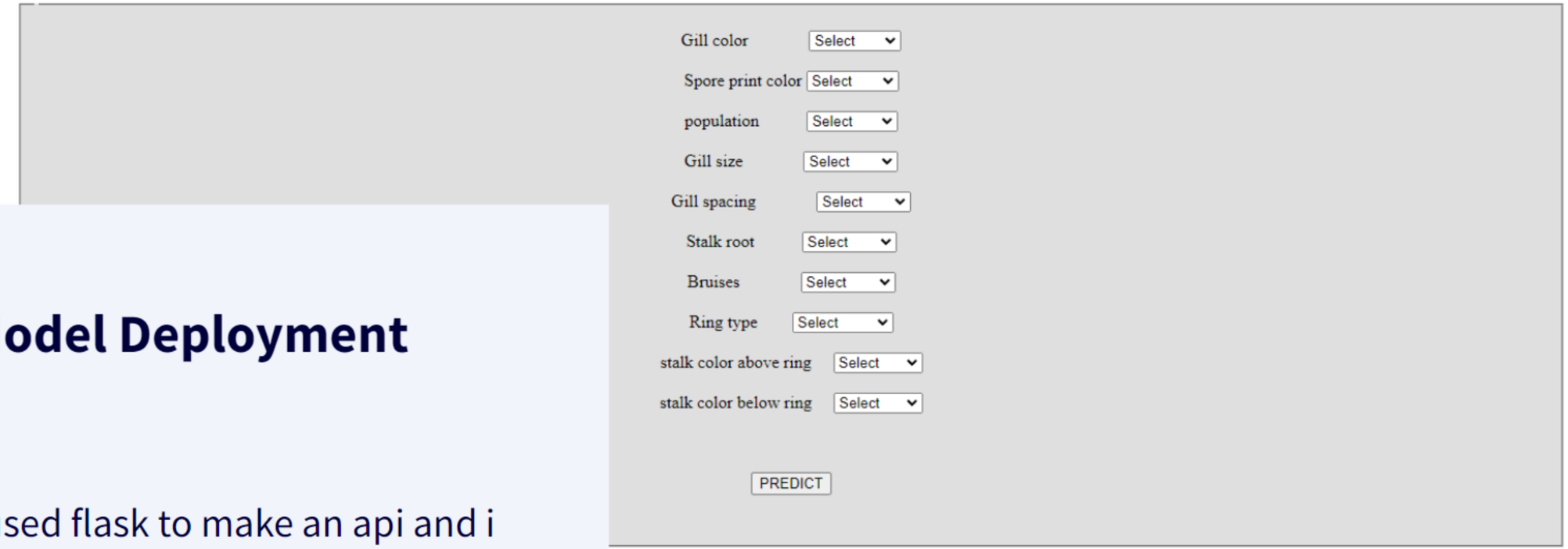
Feature Selection

In order to make it easy for the user to input his variables and get the classification I have chosen the best 10 features in the dataset with the SelectKbest Algorithm

The performance of the model dropped out alittle bit but not in a noticable way now the user can only input ten variables and get a predection easily

Attribute	Score
spore-print-color	5713.529406
ring-type	1950.610146
gill-size	1636.606833
gill-color	1401.345160
stalk-root	1186.029221
bruises	849.174461
gill-spacing	826.795274
stalk-color-above-ring	477.500664
stalk-color-below-ring	436.232423
population	311.766736

Mushroom classifier



Model Deployment

I used flask to make an api and i dumbed the model with pickle into the api

Thanks!