Classification of Mushrooms (Edible/Poisonous)

Artificial Intelligence
CS-617B

Curious 4 AI

Sacred Heart University
School of Computer Science & Engineering
The Jack Welch College of Business & Technology

Submitted To:
Dr. Reza Sadeghi

Fall - 2022

Table of Contents

Classification of Mushrooms (Edible/Poisonous)

Team: **Curious 4 AI**

Name	Contact
Sarvani Konda (Lead)	kondas2@mail.sacredheart.edu
Venkata Prasanth Pinaka	pinakav@mail.sacredheart.edu
Srikar Singam	singams2@mail.sacredheart.edu
Jayadev Varma Sri Kakarlapudi	srikakarlapudij@mail.sacredheart.edu

Introduction Of Team Members:

Sarvani Konda:

I have done my undergraduate in Information Technology and started my career in 2016. I've worked in the technology sector for over 5 years, throughout my career I have worked on Openstack, Kubernetes, Python, Golang, and other technologies related to cloud. I am an enthusiastic coder and enjoy using my skills to contribute to open-source projects. In my spare time, I love to travel, watch TV series and learn new skills.

Venkata Prasanth Pinaka:

I have completed my graduation in the stream of Computer Science and Engineering. I worked over 6 years as a developer in technologies like Oracle PLSQL and Spring boot in JAVA in a few prime organizations. I like collaborating with my team and solving the problem at hand. I enjoy researching topics and gaining knowledge.

Srikar Singam:

Enthusiastic and Driven engineer to learn and explore new things. Graduate student in Computer science at Sacred Heart University.

Jayadev Varma Sri Kakarlapudi:

I am Jayadev Varma Sri Kakarlapudi, undergraduate in Computer science engineering and worked with ADP as a software developer for almost 4 years. My interest to learn and explore more in my field of work led me here to pursue my graduation. I would like to work with people with great passion and dedication towards their work just like the bunch of people in this team.

Why we chose Sarvani Konda as team lead:

Sarvani has a proactive and enthusiastic personality. Her calm, collected and problem solving nature along with her acute knowledge on AI has led us to choose her as our Team lead. She is a teamplayer and we trust her to take the right steps for our project.

1.0 Introduction

1.1 Mushrooms:

Mushroom is the fleshy body of a fungus, typically produced above ground, on soil, or on its food source. The terms "mushroom" and "toadstool" go back centuries and were never precisely defined.During the 15th and 16th centuries, the terms mushroom is used.Identifying mushrooms requires a basic understanding of their macroscopic structure, While modern identification of mushrooms is quickly becoming molecular, the standard methods for identification are still used by most and have developed into a fine art harking back to medieval times and the Victorian era, combined with microscopic examination. Many species of mushrooms seemingly appear overnight, growing or expanding rapidly. Raw brown mushrooms are 92% water, 4% carbohydrates, 2% protein and less than 1% fat. They have minimal or no vitamin C and sodium content. Mushrooms can be used for dyeing wool and other natural fibers.Some mushrooms are used in folk medicine.

1.2 Research Question:

To classify mushrooms (North American Origin) into edible or poisonous in an easier way different machine-learning/AI models can be used. In this project, we will examine the data and build different ML/AI models that will detect the edibility of a mushroom into edible or poisonous by its specifications like gill color, cap shape, etc using classifiers.

1.3 GitHub Repository:

Here is the link for our project GitHub repository.

2.0 Dataset Description:

2.1 URL of the Dataset:

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy. [1]

2.2 When, Where and How the Dataset is collected:

These Mushroom records were taken in 1981 by G. H. Lincoff (Pres.).

Origin: Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf

Donor: Jeff Schlimmer (Jeffrey.Schlimmer '@' a.gp.cs.cmu.edu)

2.3 Name, Definition and characteristics of features:

Here we have 8124 samples in the given dataset with 22 attributes(features). The datatype of all the features is String.

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

3.0 Related Work:

3.1 Comparison of previous works:

In the paper of Schlimmer,J.S. (1987), using Representational Adjustment where a Bayesian weighting method was used that was capable of acquiring simple ideas but while implementing the four methods he faced limitations in the concept of learning tasks which lead him to Boolean chuck learning.[2]

In `Duch W, Adamczak R, Grabczewski K (1996)` paper the problem of extracting rules from neural networks has a natural geometrical interpretation. Logical description of data is possible if input spaces are correctly separated but this may become arbitrarily accurate by increasing the number of variables but this leads to disadvantage of rules becoming very large in number.[3]

In the paper `Iba,W., Wogulis,J., & Langley,P. (1988)` given a set of observations, humans acquire concepts that organize those observations and use them in classifying future experiences. This carries search through a space of possible hierarchies.[4]

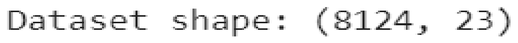
4.0 Project Plan:

By applying different Classification models on the mushroom dataset, we will be classifying mushroom edibility(edible or poisonous) . First step of the project is to clean up, transform and preprocess the data, so it can be applied on different classifier models to estimate the classification of the mushrooms.

5.0 Data Exploration:

The dataset contains 8124 rows i.e. instances of mushrooms and 23 columns i.e. the specifications like cap-shape, cap-surface, cap-color, bruises, odor, gill-size, etc.

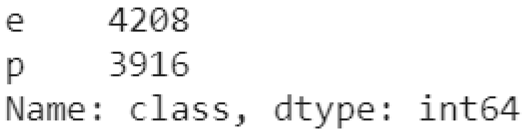
```
print("Dataset shape:", df.shape)
```



```
Dataset shape: (8124, 23)
```

The value_counts() method gives the count of the unique occurrences.

Output:



```
e    4208
p    3916
Name: class, dtype: int64
```

As we can see, there are 4208 occurrences of edible mushrooms and 3916 occurrences of poisonous mushrooms in the dataset, implying that the data is balanced.

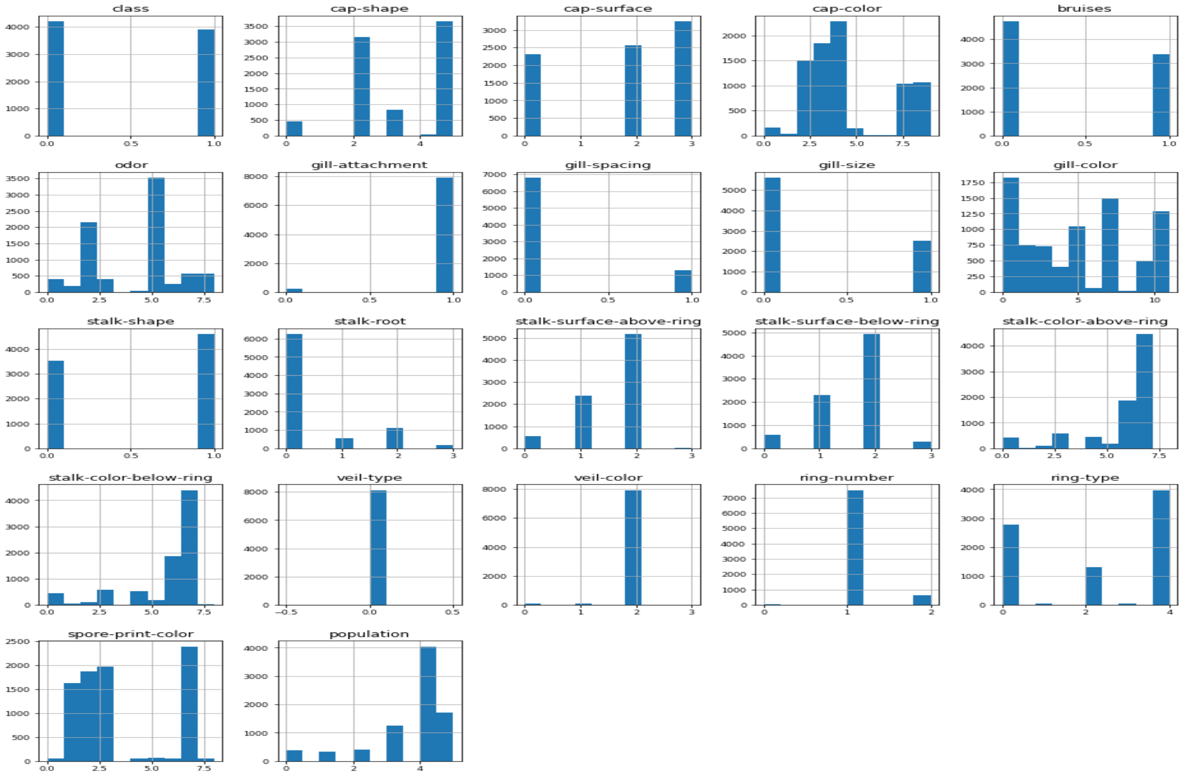
The next step is to check for missing, duplicate and null values from the selected dataset (mushrooms) and use strategies like imputation and interpolation to fill the missing data.

From our research, it is explored that there are some missing values for the feature “stalk-root”, (missing data represented by “?”). The missing values were replaced by performing an imputation strategy and estimated the value to be of “b” as it had the most frequency.

After data cleaning, the next step is to transform the data. We have used label-encoder to transform the feature values from string to integer as Scikit-learn's algorithms cannot be powered by objects or string values. It must be either in float or in integer which the Label-Encoder method provides. Next step is to analyze the data.

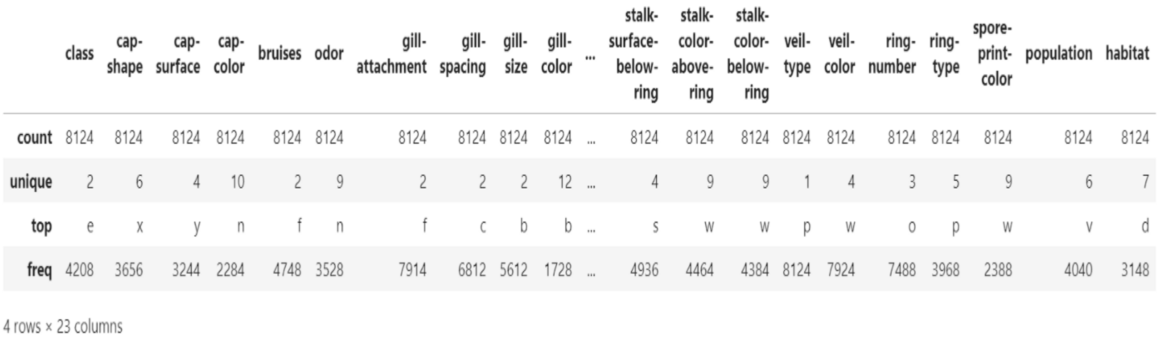
5.1 Data Analysis:

- **Univariate Analysis:** It provides the summary statistics for each feature in the data set In this project, Histogram is used on each feature of the mushroom dataset to understand and analyze the data. It can be seen from the below histograms that the features "cap-color" and "gill-color" the data is left - skewed. The feature "Veil-type" data is constant zeros.



Histogram plots on each feature

  - ○ **Descriptive Analysis**: As the name suggests, the main goal of descriptive analysis is to describe and summarize the data, different types of measures like frequency, central tendency, dispersion and etc of the dataset can be known using this method. From the describe() table in the jupyter notebook we found that the feature "stalk-color-above-ring" has the highest mean of 5.816.



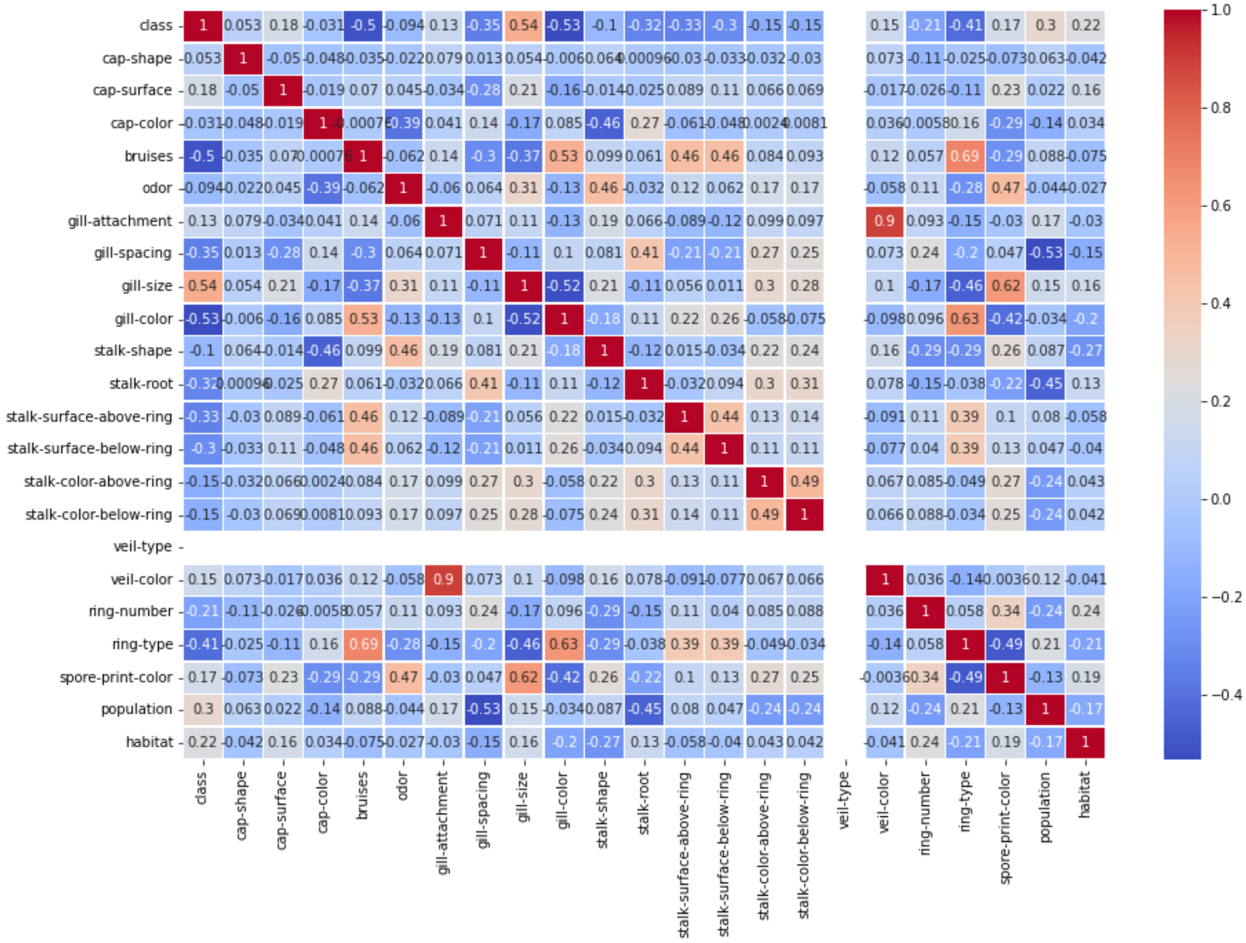
	class	cap-shape	cap-surface	cap-color	bruises	odor	gill-attachment	gill-spacing	gill-size	gill-color	...	stalk-surface-below-ring	stalk-color-above-ring	stalk-color-below-ring	veil-type	veil-color	ring-number	ring-type	spore-print-color	population	habitat
count	8124	8124	8124	8124	8124	8124	8124	8124	8124	8124	...	8124	8124	8124	8124	8124	8124	8124	8124	8124	8124
unique	2	6	4	10	2	9	2	2	2	12	...	4	9	9	1	4	3	5	9	6	7
top	e	x	y	n	f	n	f	c	b	b	...	s	w	w	p	w	o	p	w	v	d
freq	4208	3656	3244	2284	4748	3528	7914	6812	5612	1728	...	4936	4464	4384	8124	7924	7488	3968	2388	4040	3148

4 rows × 23 columns

- **Distribution Analysis**: In distribution analysis, using different statistics, graphs and density estimates of data, which will help us to identify whether the distribution of the data is a normalized distribution or if the data is skewed. It can be observed that the mushroom dataset is skewed, as histogram plots shown above are asymmetric and in the above boxplot outliers are detected.

- **Bivariate Analysis:** This is another method to analyze more of the dataset by finding the relationship between each feature of the dataset and the target variable or a relationship between any 2 features. Ex: Pair-plot, Scatter plot, etc.

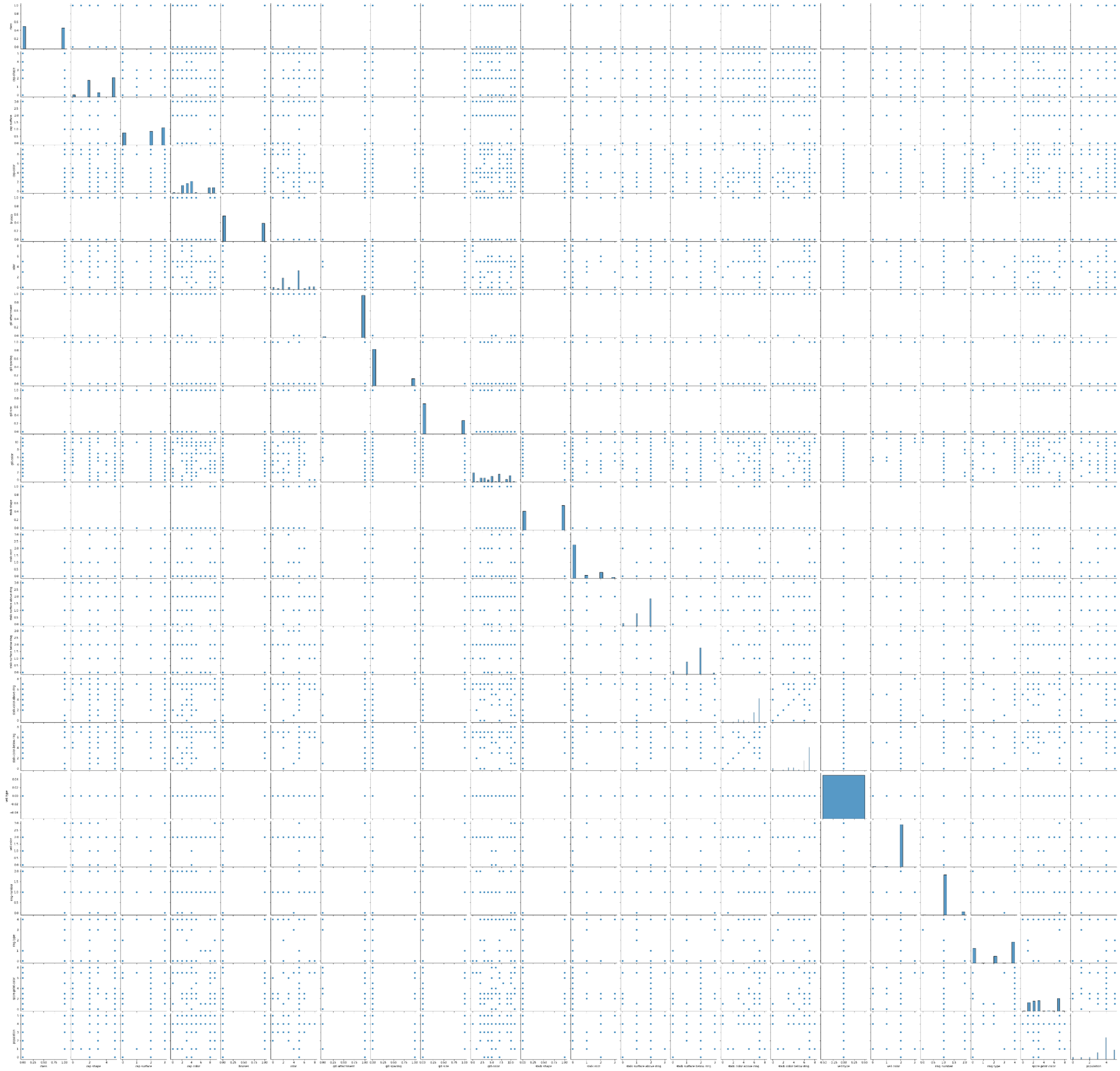
  - **Pearson Correlation:** It is a method for numerical variables, it assigns a value between -1 to 1 for all the features, where 0 signifies no relationship, 1 is total positive correlation, and − 1 is total negative correlation. A correlation value of 1 between two variables would indicate that a significant and positive relationship exists between the two.



HeatMap

From the above heatmap plot, it is observed that the feature veil-type is assigned a 0 value indicating that there is no relationship between veil-type and any other feature, and veil-color is highly correlated with the feature gill-attachment with the correlation value of a positive 0.9 which means if veil-color increases gill-attachment also increases.

○ **Pair Plot:** It allows us to understand the linear and non-linear relationships among features. By the below Pair plot it is observed that cap-color and gill color are non linearly related when compared to other features.


class
cap-shape
cap-surface
cap-color
bruises
odor
gill-attachment
gill-spacing
gill-size
gill-color
stalk-shape
stalk-root
stalk-surface-above-ring
stalk-surface-below-ring
stalk-color-above-ring
stalk-color-below-ring
veil-type
veil-color
ring-number
ring-type
spore-print-color
population

Pair Plot

6. Data Modeling:

In the data modeling phase, machine learning Algorithms are applied on the mushroom dataset and are trained, tuned and validate a ML model.

6.1 Preprocessing:

It is observed that the given data has a zig zag bell curve distribution, to maintain the curve uniformity the data is standardized using the StandardScaler().

6.2 Data splitting:

Setting X and y-axis and splitting the data into train and test respectively. Since we want to predict the class of the mushroom, we will drop the ‘class’ column.

The entire mushroom dataset is divided into train and test sets in 80 : 20. The 80 percent of the data is used to train the model and 20 percent is used for validation and is divided as 6499 and 1625 samples.

```
X = df.drop([‘class’], axis=1)
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,rando
m_state=42,shuffle=True, stratify=y)
print('X_train size:',X_train.shape)
print('X_test size:',X_test.shape)
```

Train and Test sample count for mushroom dataset:

```
X_train size: (6499, 21)
X  test size: (1625, 21)
```

6.3 Fitting the model:

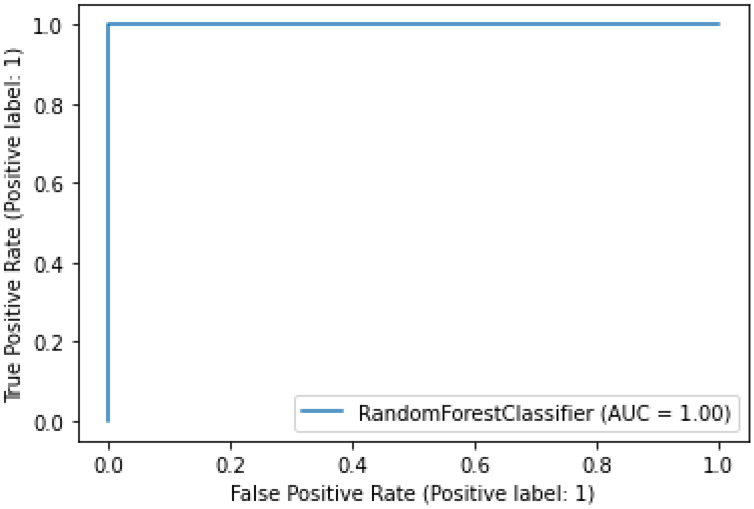
In this step, Three classifiers models (Random Forest, K Nearest Neighbors, Decision Tree) are applied on the mushroom dataset to predict the classification of the mushroom.

6.4 Measuring Performance:

Importing accuracy score, classification report, confusion matrix, plot_roc curve.

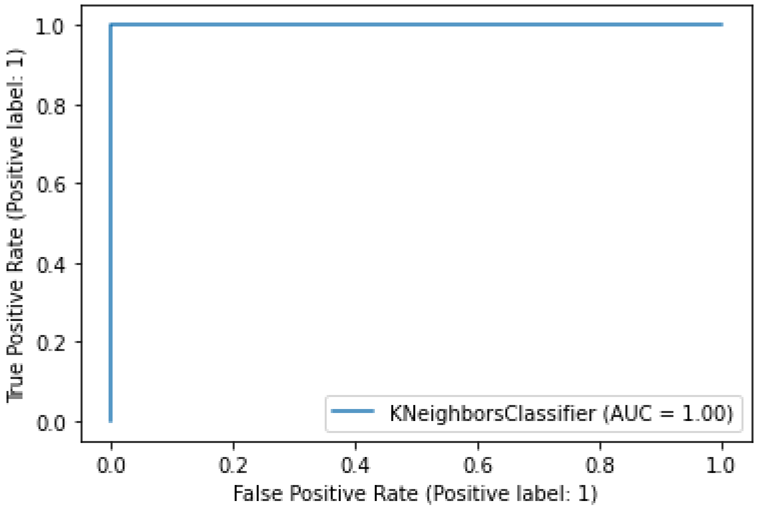
Prediction of Accuracy, Confusion Matrix, ROC Curve: Using Random Forest model:

```
The accuracy of RF 1.0
RF model details
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       842
           1       1.00      1.00      1.00       783

    accuracy                           1.00      1625
   macro avg       1.00      1.00      1.00      1625
weighted avg       1.00      1.00      1.00      1625

TN = 842, FP = 0, FN = 0, TP= 783
```



Prediction of Accuracy, Confusion Matrix, ROC Curve: Using KNN (K Nearest Neighbors):

```
The accuracy of KNN 1.0
KNN model details
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       842
           1       1.00      1.00      1.00       783

    accuracy                           1.00      1625
   macro avg       1.00      1.00      1.00      1625
weighted avg       1.00      1.00      1.00      1625

TN = 842, FP = 0, FN = 0, TP= 783
```



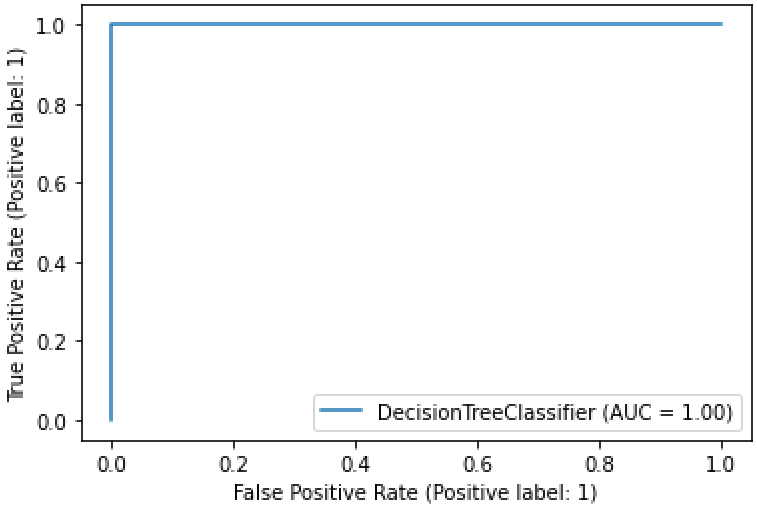
Prediction of Accuracy, Confusion Matrix, ROC Curve: Using Decision Tree:

```
The accuracy of DT is 1.0
DT model details
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       842
           1       1.00      1.00      1.00       783
```

```
    accuracy                           1.00      1625
   macro avg       1.00      1.00      1.00      1625
weighted avg       1.00      1.00      1.00      1625
```

TN = 842, FP = 0, FN = 0, TP= 783

```
<sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x7f45860b4690>
```



From the above confusion matrices of all the three models it is observed that there are 824 True Positives and 783 True negatives cases are correctly predicted.

Therefore, the accuracy is 1.00 stating that the model has a correct prediction percent of 100%.

F1-Score = 2* precision * recall/ precision + recall = 1.00.

The calculated values and the classification report metrics are thus verified. The ROC for the model is plotted and the obtained AUC value is 1.00 indicating the classifier model is able to distinguish the positive class values from the negative class values correctly.

All the three models Random Forest, K Nearest Neighbors, Decision Tree have the accuracy and f1 score as 1.00 implying that the models have predicted the classification of mushrooms correctly.

7. Optimization and Model Evaluation:

Though the accuracy of the models used for prediction is 100% for the mushrooms dataset we are trying to see if there is any difference in feature importance while using optimization techniques.

We are using K-Fold splitting strategy(k=8), RandomizedSearchCV as hyper parametric optimizer and Cross_validation as model validation.

Steps for cross-validation:

- Dataset is split into K "folds" of equal size.
- Each fold acts as the testing set 1 time, and acts as the training set K-1 times.
- Average testing performance is used as the estimate of out-of-sample performance also known as cross-validated performance.

DT optimization:

Results:

Accuracy mean 99.19%

Accuracy interval 97.44% 100.93%

Parameters:

criterion: gini

max_depth: 30

min_samples_leaf: 45

min_samples_split: 114

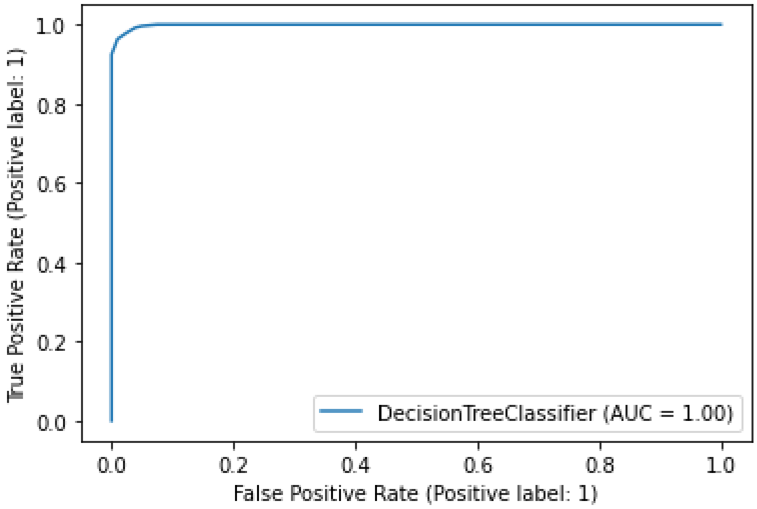
The accuracy of D_T is 0.9772307692307692

D_T model details

```
          precision    recall  f1-score   support

       0       0.98      0.98      0.98       842
       1       0.97      0.98      0.98       783

    accuracy                           0.98      1625
   macro avg       0.98      0.98      0.98      1625
weighted avg       0.98      0.98      0.98      1625
```

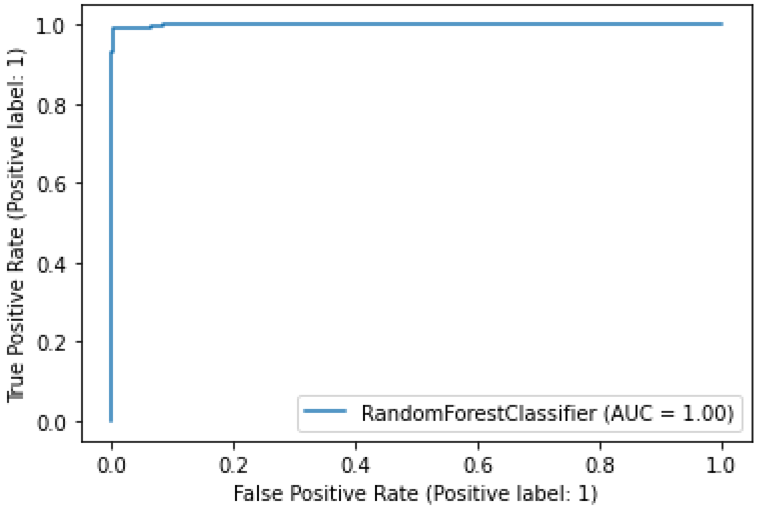
TN = 822, FP = 20, FN = 17, TP= 766

```
<sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x7f36ab0c1970>
```



KNN optimization:

```
Results:
Accuracy mean 99.89%
Accuracy interval 99.62% 100.16%


Parameters:
weights: distance
n_neighbors: 7
algorithm: auto
The accuracy of KNN is 1.0
KNN model details
              precision    recall  f1-score   support


           0       1.00      1.00      1.00       842
           1       1.00      1.00      1.00       783


    accuracy                           1.00      1625
   macro avg       1.00      1.00      1.00      1625
weighted avg       1.00      1.00      1.00      1625


TN = 842, FP = 0, FN = 0, TP= 783
<sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x7f36aaf60a60>
```



RF optimization:

Results:

Accuracy mean 99.00%

Accuracy interval 98.66% 99.35%

Parameters:

n_estimators: 48

min_samples_split: 5

min_samples_leaf: 1

max_features: auto

max_depth: 4

bootstrap: True

The accuracy of RF is 0.9883076923076923

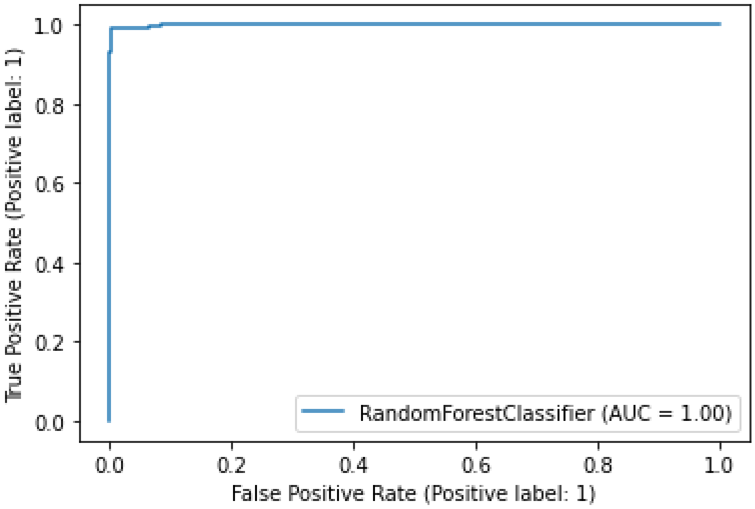
RF model details

```
          precision    recall  f1-score   support
```

```
       0      0.98      1.00      0.99      842

       1      1.00      0.98      0.99      783


    accuracy                        0.99      1625

   macro avg      0.99      0.99      0.99      1625

weighted avg      0.99      0.99      0.99      1625
```

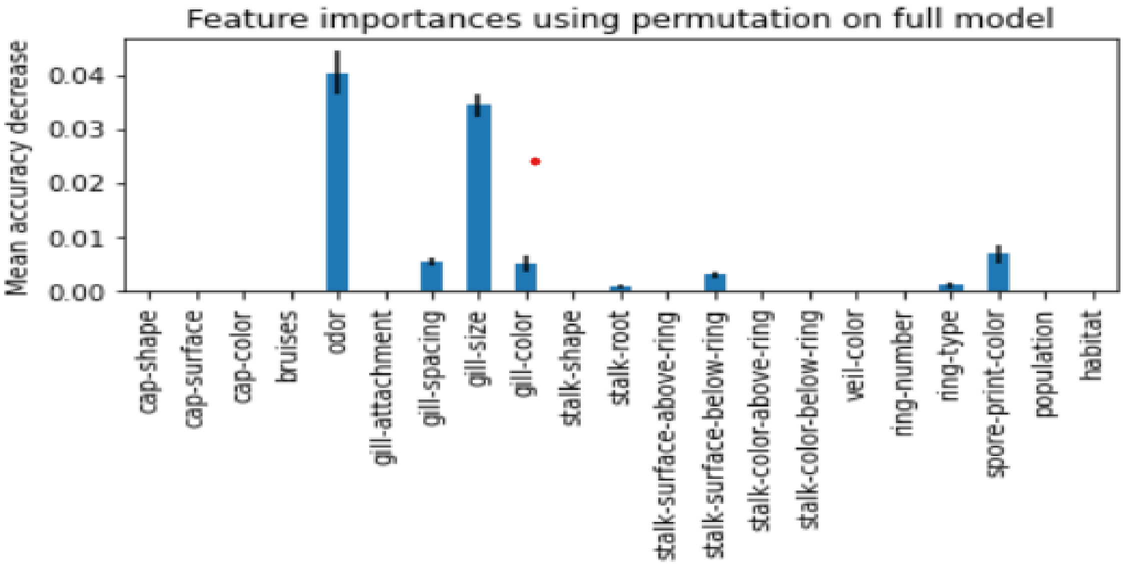
TN = 839, FP = 3, FN = 16, TP= 767

```
<sklearn.metrics._plot.roc_curve.RocCurveDisplay at 0x7f36aae70fa0>
```



Feature Importance:

From the below feature importance using permutation image it can be observed that the ‘Odor’ feature has more importance than the rest, followed by ‘Gill Size’.



8.0 Conclusion:

RF, KNN and DT are studied and experimented on the mushroom(North American mushrooms) dataset to classify the mushrooms into edible or poisonous.

The dataset exploration has proven that the data is skewed for all features and the target is balanced.

Even though all considered models have shown 100% accuracy before optimization after optimizing using 'Randomized search cross validation' the accuracy was slightly reduced to 99.3, 98.9 and 99.9 for Decision Tree, Random Forest and KNN respectively. As KNN has the best accuracy amongst the three models thus it's a better model for the given Mushrooms dataset.

The feature importance using permutation for RF has shown that the 'Odor' feature has more importance than the rest, followed by 'Gill Size'.

9. References:

1. https://archive.ics.uci.edu/ml/datasets/mushroom
2. https://escholarship.org/uc/item/48r6d4z0
3. https://www.researchgate.net/publication/2633347_Extraction_of_Logical_Rules_From_Training_Data_Using_Backpropagation_Networks
4. https://www.sciencedirect.com/science/article/abs/pii/0004370289900465