Testing The Edibility of Mushrooms Using Machine Learning Algorithms and Data Analysis

Tilottoma Barua

Contents

1. Introduction:

The project is based on mushroom's data samples. Based on the dataset, the mushrooms are classified as either edible or poisonous. This dataset is available in UCI data repository which records all Agaricus and Lepiota family of mushrooms. These mushrooms are classified as edible or poisonous depending upon various features. This is used as a training dataset for my proposed project. In this project, different types of machine learning algorithms have been used to model the dataset and used that model to predict the edibility of the mushrooms. Also, relevant data analysis techniques have been used with proper graph representations to find out the most important features which have the maximum information gain to differentiate between edible and poisonous mushrooms.

2. Data Sets:

A well-maintained dataset with adequate features on mushroom species is required for this project. Fortunately, an organized dataset is found at the following URL which is used with a proper acknowledgment:
https://www.kaggle.com/uciml/mushroom-classification/data

2.1. Dataset File Description:

Target variable:

- class: e=edible, p=poisonous

Features:

- 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: bruises=t, no=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, missing=?
- 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

3. Hypothesis:

I. I want to find the most important features which have the maximum information gain to differentiate between edible and poisonous mushrooms. Also, I want to build a predictive model to predict the edibility (which is eligible to eat) of mushrooms using ML algorithms. This model will predict the edibility of the mushrooms using the various features with considerable accuracy.

II. I assume that odor, gill color and habitat (**null hypo 1**) could be the most important features to differentiate between edible and poisonous mushrooms. A foul, pungent odor **(null hypo 2)** and black gill color (**null hypo 3)** on the mushroom could be the indication for poisonous mushroom.

III. On the other hand, broad gill size **(alt hypo 1**) and orange spore print color (**alt hypo 2**) could be another feature of edible mushroom. Solitary population **(alt hypo 3**) could be another feature to distinguish between edible and poisonous mushrooms.

IV. I Found a dataset which has more than 15 features of mushrooms with edible and poisonous class labels. This dataset can be used to find the most important features which have the maximum information gain. Random forest's feature importance method can be used to find important features.

4. Test Plan:

To test the above hypothesis, several tests have been run in the following manners:

I. Load into the data frames from the .csv file to describe the data set and draw a pie plot to check whether the target dataset is balanced or not.

II. Split the dataset into the train and test data set and apply several Machine Learning Algorithms to model it and compare their accuracy.

III. Vary the depth of the decision tree algorithm to check whether accuracy reduces with the shallowness of the tree. That means the shallower the decision tree is, the less accurate the model will be.

IV. Find out the most important features which have maximum information in differentiating between edible and poisonous mushrooms.

V. Re-train the ML models to check the accuracy level with the best 5 features obtained from step IV.

VI. Check the null-hypothesis by using relevant graphs and bar plots and argue which hypothesis is correct and which is wrong.

VII. Check the alt-hypothesis by using relevant graphs and bar plots and argue which hypothesis is correct and which is wrong.

5. Results and Analysis:

5.1. Test Plan I:

Some Basic Info of the Dataset

	class	cap-shape	cap-surface	...	spore-print-color	population	habitat
count	8124	8124	8124	...	8124	8124	8124
unique	2	6	4	...	9	6	7
top	e	x	y	...	w	v	d
freq	4208	3656	3244	...	2388	4040	3148

[4 rows x 23 columns]

Fig 1: Some basic info of the data set

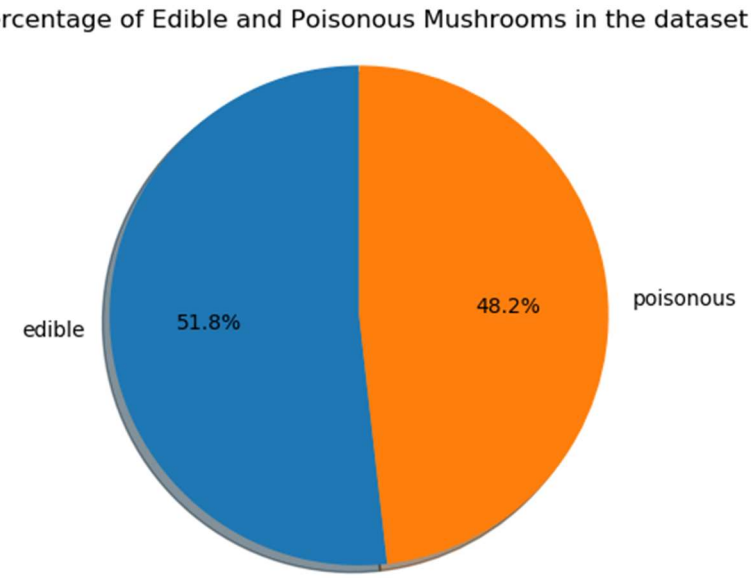


Fig 2: Pie chart showing the percentage of edible and poisonous mushroom in the dataset

5.2. Result Analysis of Test Plan I:

It is learned from the basic info of the dataset that this dataset has 8124 instances with 23 features. The features are presented using some form of shorthand notation (e.g. e = edible and p = poisonous). This short hand alphabetic notation must be converted into unique integers to use them in the ML algorithms. So. label encoding has been performed on the data. And we have avoided one hot encoding because it is not suitable for this dataset.

The pie chart shows the percentage of poisonous and edible mushrooms in the overall dataset. We can see from the pie plot is that both have almost equal percentage (edible-51.8% and Poisonous-48.2%) in the overall dataset. This indicates that this dataset is balanced in terms of targets classes and the chance of having a biased result is low.

As the dataset has a total of 23 features, we can use these features to build a Machine learning predictive model to predict the edibility (which is eligible or not) of mushrooms. Here 'class' is our target variable and the remaining 22 columns are our features for the ML models.

5.3. Test Plan II:

Fig 3: Classification Report and Heat map for Logistic Regression ML Algorithm

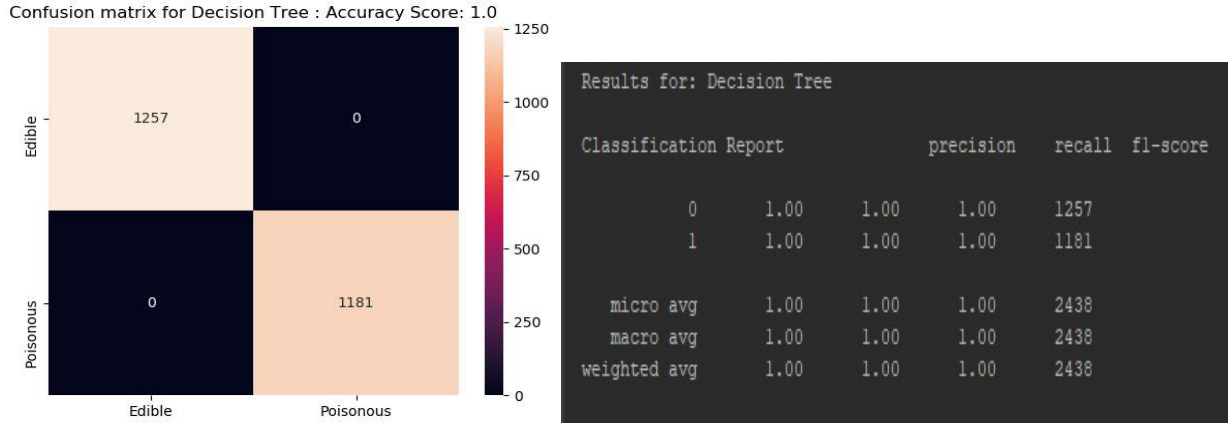
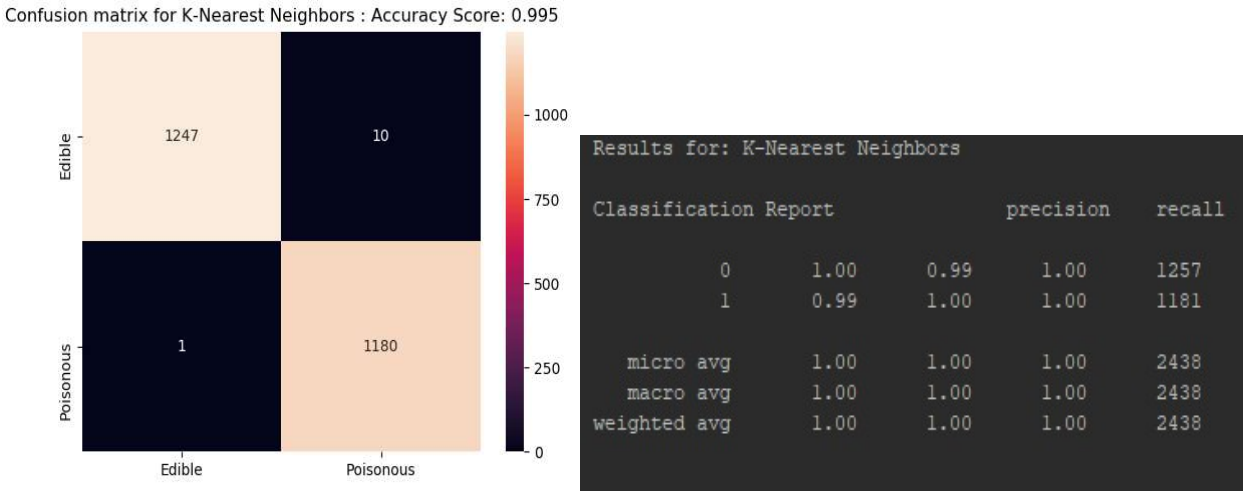


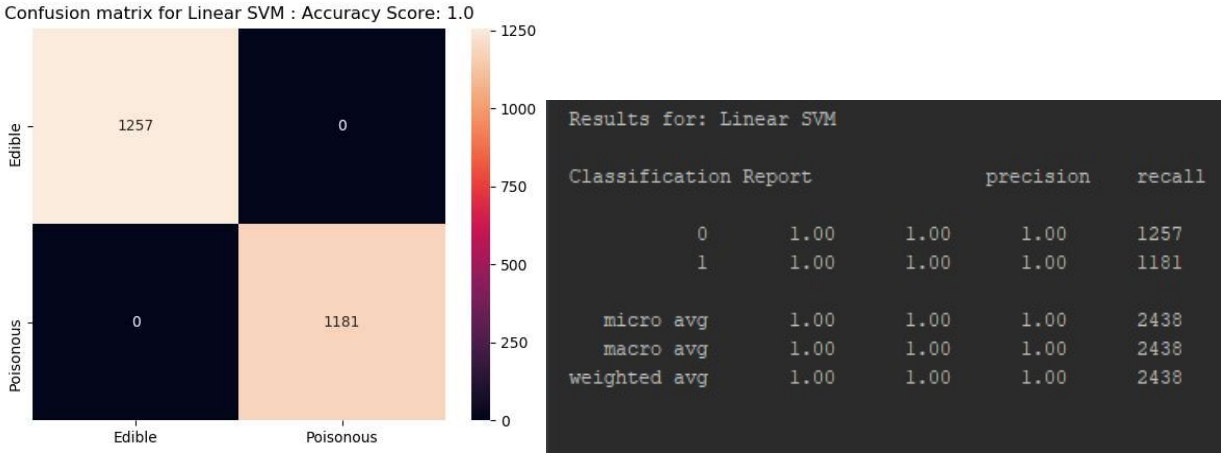
Fig 4: Classification Report and Heat map for Decision Tree ML Algorithm



Results for: K-Nearest Neighbors

Classification Report			precision	recall
0	1.00	0.99	1.00	1257
1	0.99	1.00	1.00	1181
micro avg	1.00	1.00	1.00	2438
macro avg	1.00	1.00	1.00	2438
weighted avg	1.00	1.00	1.00	2438

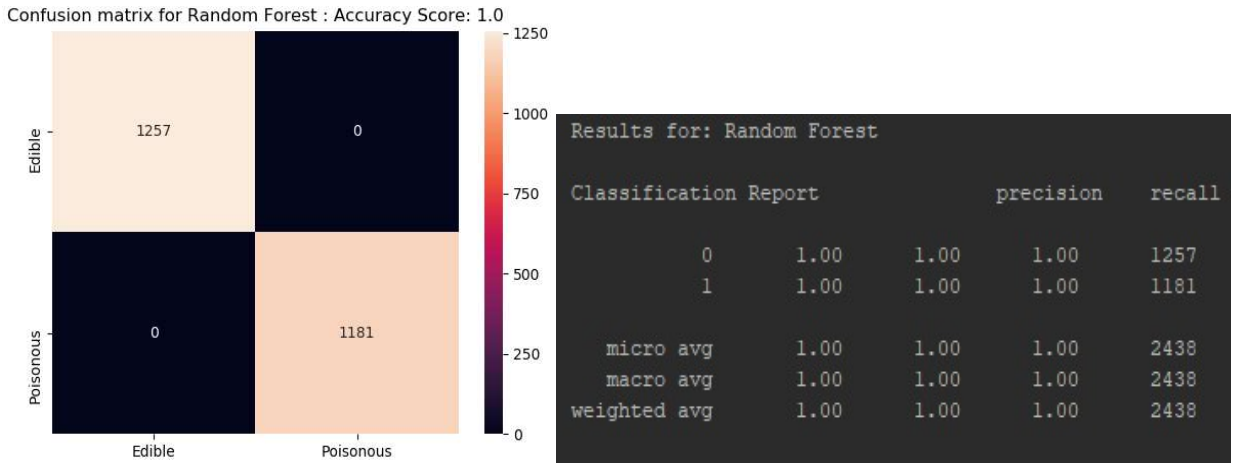
Fig 5: Classification Report and Heat map for K-Nearest Neighbor ML Algorithm



Results for: Linear SVM

Classification Report			precision	recall
0	1.00	1.00	1.00	1257
1	1.00	1.00	1.00	1181
micro avg	1.00	1.00	1.00	2438
macro avg	1.00	1.00	1.00	2438
weighted avg	1.00	1.00	1.00	2438

Fig 6: Classification Report and Heat map for Linear SVM ML Algorithm



Results for: Random Forest

Classification Report			precision	recall
0	1.00	1.00	1.00	1257
1	1.00	1.00	1.00	1181
micro avg	1.00	1.00	1.00	2438
macro avg	1.00	1.00	1.00	2438
weighted avg	1.00	1.00	1.00	2438

Fig 7: Classification Report and Heat map for random Forest ML Algorithm

5.4. Result Analysis of Test Plan II:

For this test part, I have used several machine learning algorithms to model the dataset. At first, this large dataset is split into test and training data and 5 different machine learning algorithms are used to model the dataset and after that their accuracies are compared.

a) **Logistic Regression:** This model is almost able to correctly predict the edibility of the mushroom and achieved about 94.9% accuracy score.

b) **Decision tree Algorithm, Random Forest & SVM Linear Algorithm**: These models gave the perfect result to predict the edibility of the mushroom and achieved 100% accuracy score. As per the result, these model fully supports the hypothesis and our requirement. It seems that, all the tree-based methods and the linear SVM all achieved 100% accuracy but there could be a chance that the data is overfitting. However, in the next test, we will analyze whether the accuracy varies with the tree depth or not.

c) **K- Nearest Neighbor:** This machine learning model gave also a very good prediction model with a very good score of about 99.5% accuracy. So, we can also consider KNN as a usable algorithm to model this dataset.

5.5. Test Plan III:

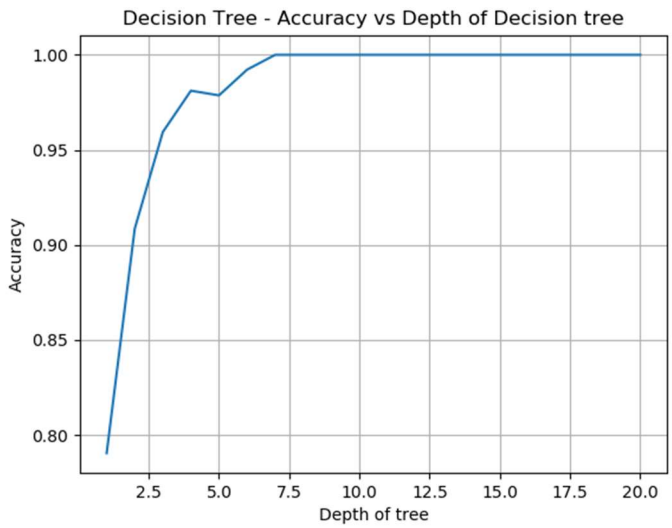
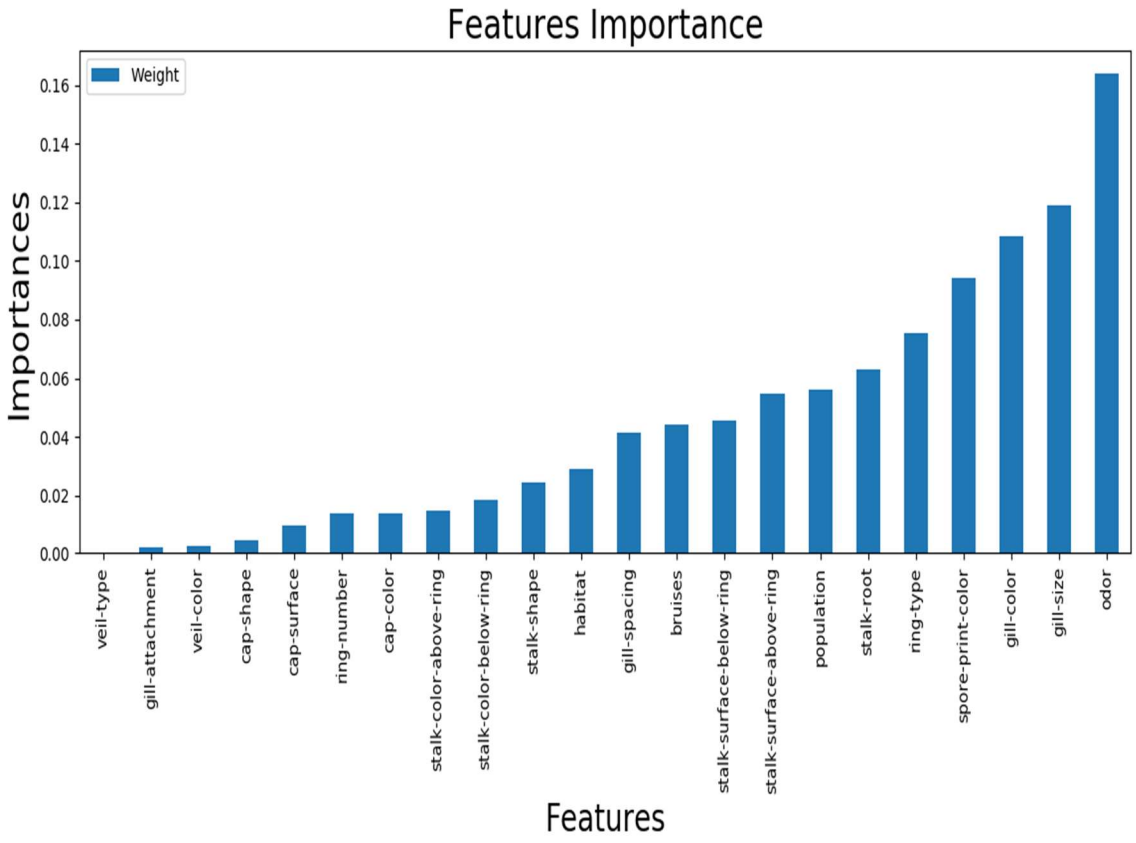


Fig 8: Decision Tree- Accuracy vs Depth of the tree graph

5.6. Result Analysis of Test Plan III:

As all the tree models give 100% accuracy, so there could be a chance of overfitting. I have used the depth of the decision tree to check whether accuracy varies with the decision tree depth. From Fig. 8, it is clear that decision tree model has achieved its highest accuracy (100%) at 7^{th} depth. Even for very low depth (e.g. 1 or 2) the accuracy is close to 90%. It could be a reason that the features of the mushroom dataset are orthogonal. And also these features are properly picked and have high information about the edible and poisonous nature of the mushrooms. This could be a reason for getting high accuracy even for shallower decision tree.

5.7. Test Plan IV:



Sorted Feature Importance

	Features	Weight
15	veil-type	0.000000
5	gill-attachment	0.001996
16	veil-color	0.002488
0	cap-shape	0.004927
1	cap-surface	0.010001
17	ring-number	0.013698
2	cap-color	0.014095
13	stalk-color-above-ring	0.014904
14	stalk-color-below-ring	0.018305
9	stalk-shape	0.024394
21	habitat	0.029048
6	gill-spacing	0.041422
3	bruises	0.044064
12	stalk-surface-below-ring	0.045620
11	stalk-surface-above-ring	0.054947
20	population	0.055944
10	stalk-root	0.063035
18	ring-type	0.075431
19	spore-print-color	0.094184
8	gill-color	0.108559
7	gill-size	0.119143
4	odor	0.163798

Fig 9: Feature Importance

5.8.Result Analysis of Test Plan IV:

To find out the most important features in the dataset to classify between edible and poisonous mushrooms, random forest feature importance method has been used to find out the weight of all the features. Then the feature's weights are sorted according to their importance. From Fig. 9, we can find out the most important five features responsible for the edibility of the mushroom classification. They are 'Odor', 'Gill Size', 'Gill-Color', 'Spore Print Color' and 'Ring Type'.

5.9.Test Plan V:

```
************************************************
After retraining, Accuracy Score Summary
                 model  accuracy score
0  Logistic Regression        0.885972
1        Decision Tree        0.994668
2  K-Nearest Neighbors        0.994668
3           Linear SVM        0.994668
4        Random Forest        0.994668
```

Fig 10: Re-train the model with 5 most important feature

5.10. Result Analysis of Test Plan V:

Now after getting the most important features from test plan IV, we re-train our models to check the accuracies of the same five ML algorithms. From Fig. 10, we can see that, except for the Logistic regression model, all other machine learning models give about 99.8% accuracy. This shows that the five features that, we have picked can be used as the most indicative feature to find out the edible and poisonous mushrooms. From Fig. 9 and Fig. 10, we can say that we have already tested our **null hypothesis-1** that, Odor, Gill color are two of the most important features to differentiate between the edible and the poisonous mushroom. However, habitat is also a good feature but not the most important five to be considered.

5.11. Test Plan VI:

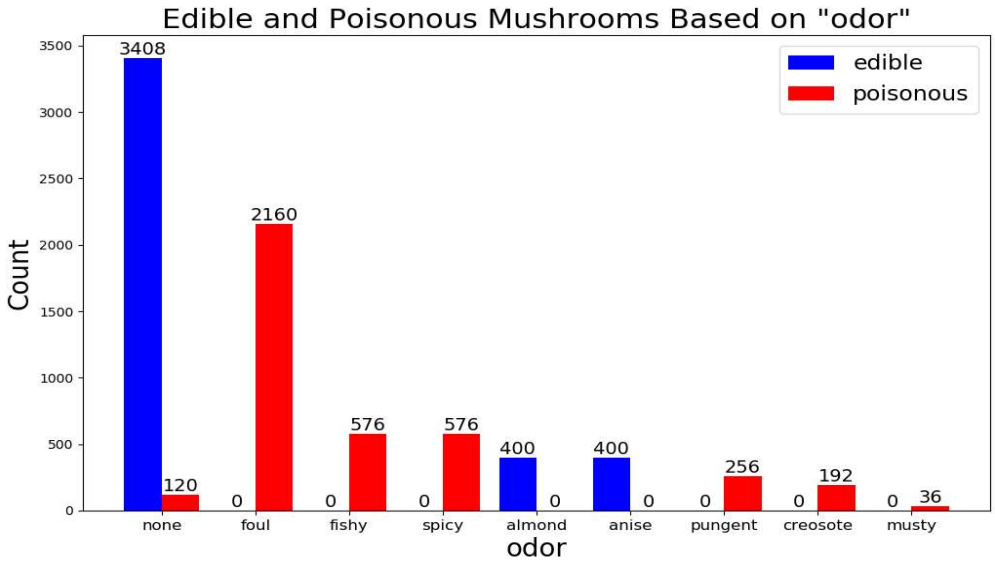


Fig 11: 'Odor' vs 'mushroom-class' Count plot

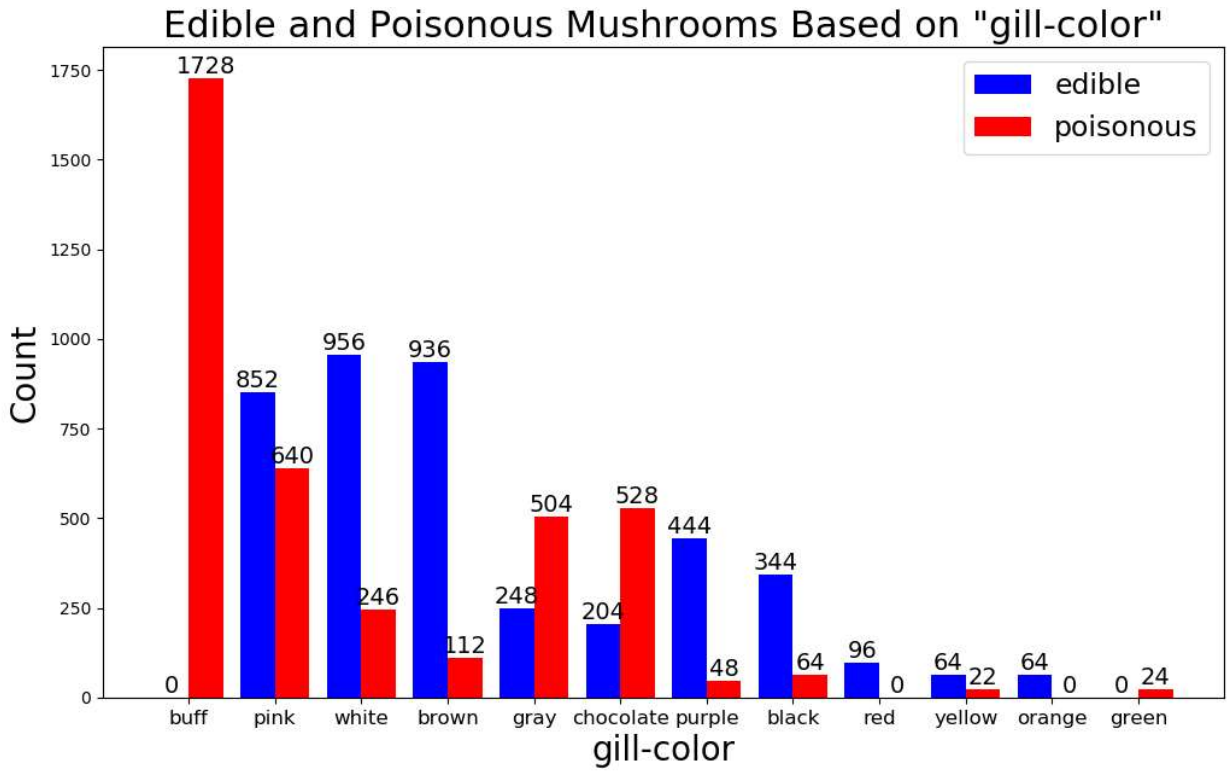


Fig. 12: 'gill-color' vs 'mushroom-class' Count plot

5.12. Result Analysis of Test Plan VI:

From Fig. 11 and Fig. 12, we can test our **null hypo 2 & null hypo 3**. From Fig. 11, we can see that, there is a total of 2160 mushrooms which have **foul odor** and 256 mushrooms which have a **pungent odor**. And all these mushrooms are poisonous. There are no such mushrooms in the dataset which are edible but have a foul and pungent odor. **So, we can conclude safely that the mushrooms which have foul and pungent odor are always poisonous**. So **null hypo 2** is tested and verified.

From Fig. 12, we can test our **null hypo 3.** From this figure, we can see that **black gill color** is one of the features for poisonous mushrooms but not all the black gill-color mushrooms are poisonous. 344 mushrooms with black gill have been proved to be edible compared to only 64 poisonous mushrooms having the same black gill color. **So, null hypo 3 is not right**. On the other hand, mushrooms having buff- gill color are always poisonous.

5.13. Test Plan VII:

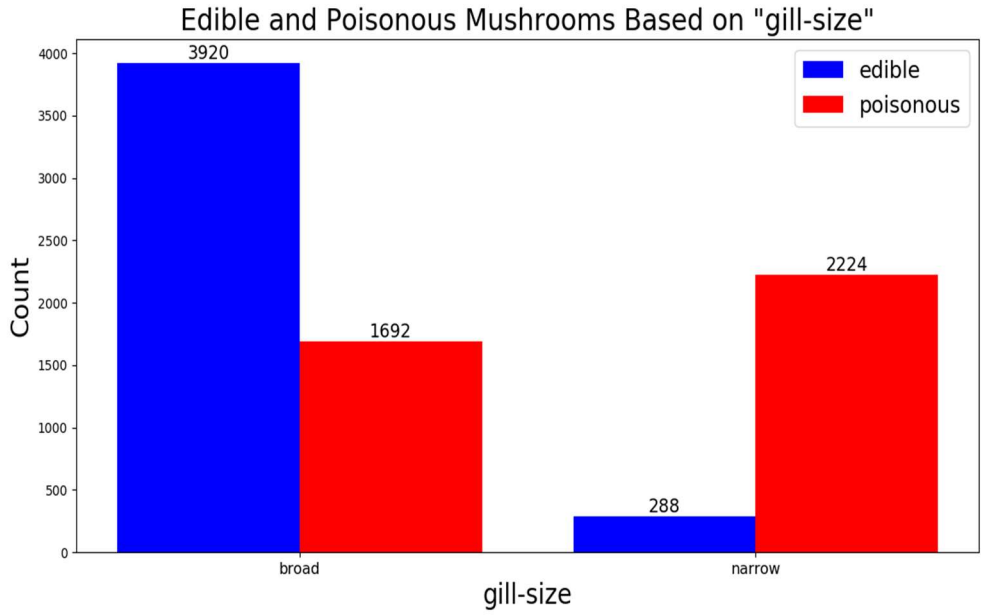


Fig 13: 'gill-size' vs 'mushroom-class' Count plot

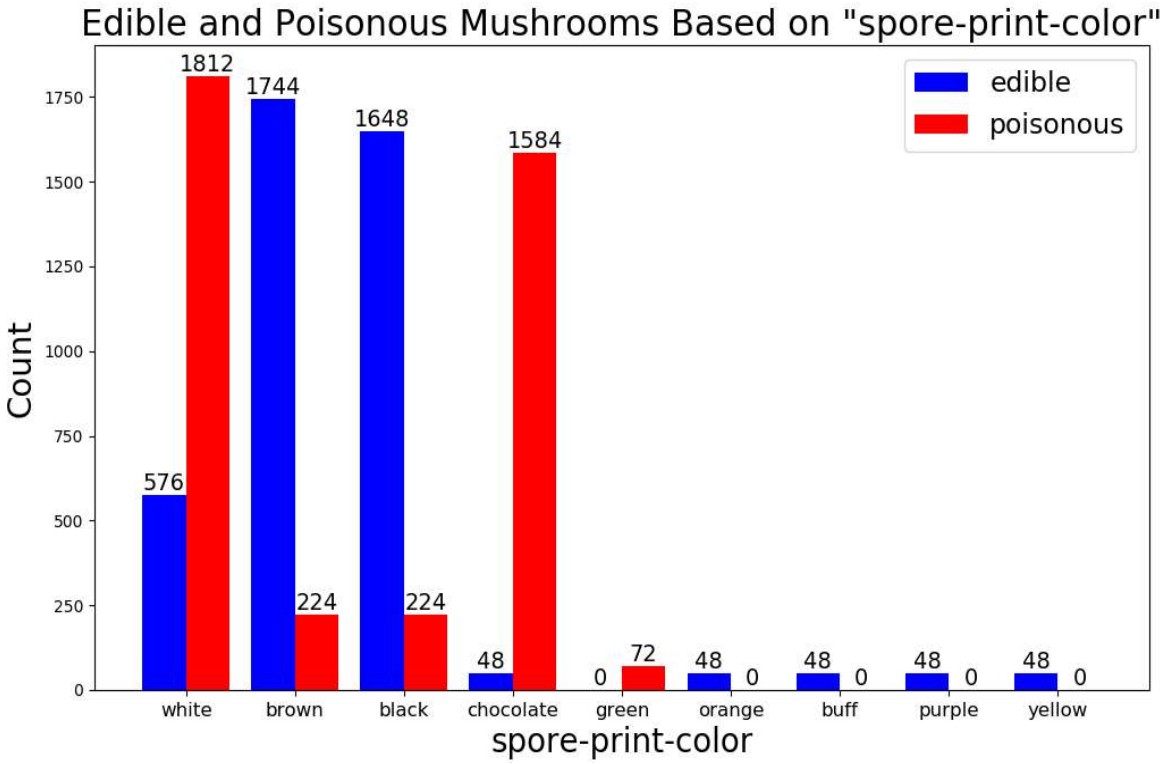


Fig 14: 'Spore Print color' vs 'mushroom-class' Count plot

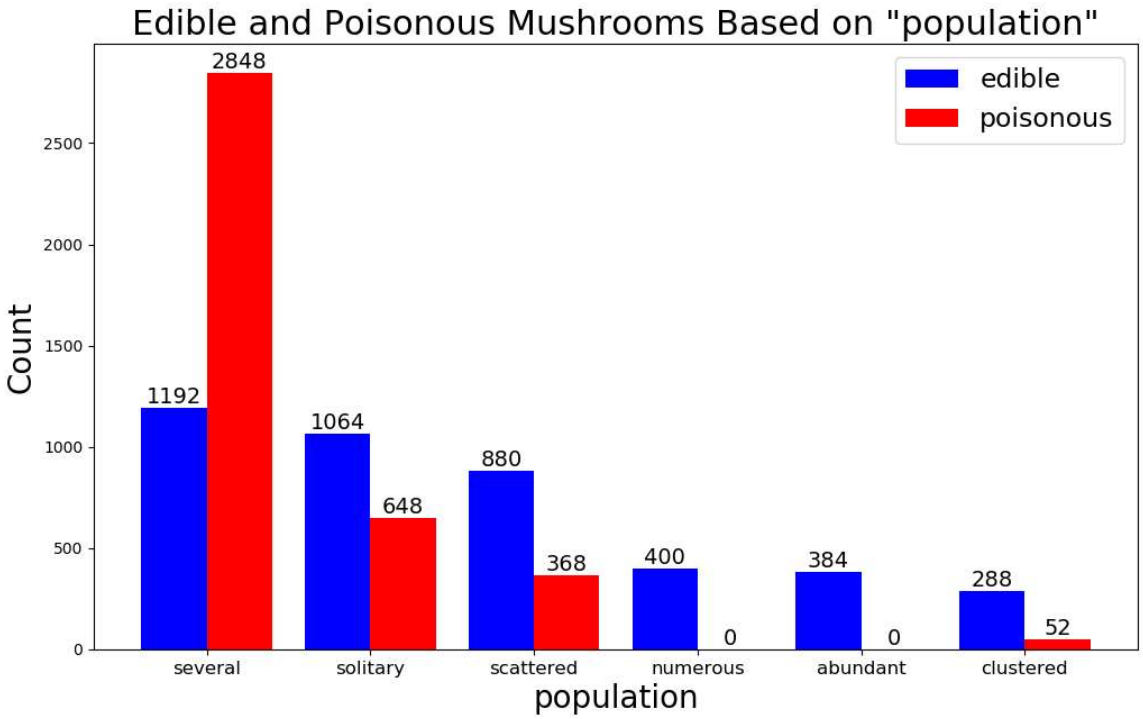


Fig 15: 'population' vs 'mushroom-class' Count plot

5.14. Result Analysis of Test Plan VII:

From Fig. 13, we can analyze **alt hypo 1**. There are 3920 edible mushrooms and 1692 poisonous mushrooms having the same broad gill size**. So, we cannot safely conclude that all the mushrooms having broad gill size are edible**. **So, our alt hypo 1 is not completely true**. Rather, a mushroom having narrow gill size has good probability to be poisonous. There are 2224 mushrooms with narrow gill are found poisonous compared to 288 edible mushrooms with narrow gill size.

We can analyze **alt hypo 2** from Fig. 14. No mushroom with orange spore print color is poisonous. All the mushrooms having orange print color is edible. **So, our alt hypo 2 is true**. However, white and chocolate spore print color is the most important features here for a mushroom to be poisonous.

We can analyze **alt hypo 3** from Fig. 15. I assumed that the solitary population could be another feature to distinguish between edible and poisonous mushrooms. Here, we can see that only 648 solitary mushrooms are found poisonous whereas, 1048 solitary are found as edible. **So, by using the only solitary population as a feature we cannot safely distinguish between poisonous and edible mushrooms**. **So, our alt hypo 3 is not true.** Rather, numerous and abundant population can be used to safely distinguish between edible and poisonous mushrooms.

6. Conclusion:

Five different ML algorithms have been used to model the dataset using 22 features and predict the edible and poisonous nature of the mushrooms from the models. Their accuracies have been listed and compared. Random forest feature importance method has been used to find the most

important features from the dataset to distinguish between edible and poisonous mushrooms. The five models have been retrained again with only five most important features obtained in the last step and their accuracies have been compared again. Relevant data analysis techniques and bar plots are used to argue and prove the null hypothesis and alternative hypothesis and results are presented in a comprehensive manner.

7. Future Work:

We have found the important features in the mushroom dataset using the Random Forest Classifier. Another way could be use 'entropy' instead of random forest importance and can be compared with the current result. Another work will be to find out the exact reason why all the five ML algorithms have given close to 100% accuracy. The reason could be overfitting. The approach will be to use K-fold cross validation to check the model accuracy.