IS 603
Decision Making Support System

Final Project Report

On

Classification Of Mushroom: “Edible” Or “Poisonous”

Submitted to: Dr. James Foulds

Submitted by: Group 7- Neha Upadhyay, Sudeep Mishra, Sumati Mane, Travez Hardin, Yash Pachorkar

Table of Contents

1. Abstract:

In this study, a classification model is applied to data in order to determine if mushrooms are toxic or edible. The data is totally nominal and categorical in nature. The data was collected as part of a Kaggle competition and is also available on the UCI Machine Learning repository. Finding the best functioning model and deriving conclusions regarding mushroom taxonomy were among the goals. Decision Tree (C4.5), NaiveBayes and Support Vector Machine (SVM), and Logistic Regression are four comparisons of the best categorization techniques in data mining. We used the Weka Tool to test the approaches, and we used Python Programming for data analysis and visualization. According to the results of the tests, the C4.5 algorithm is 100 percent as accurate as the SVM and logistic regression, whereas NaiveBayes is a little over 95 percent accurate. The C4.5 algorithm, on the other hand, is faster than both in terms of speed.

2. Introduction:

Humans have been eating mushrooms, the fruiting body of fungi, for thousands of years. Although all mushrooms include protein, fiber, and the potent antioxidant selenium, certain varieties are prized for their health benefits. Shiitake mushrooms, for example, contain all eight essential amino acids as well as eritadenine, a cholesterol-lowering substance. Reishi mushrooms are prized for their immune-boosting benefits, maitake mushrooms for their blood sugar-stabilizing properties, and porcini mushrooms for their anti-inflammatory properties. Including mushrooms in one's diet used to entail foraging, which carried the risk of consuming toxic mushrooms. Many species of mushrooms, however, have been successfully cultivated since the 1600s. Agaricus bisporus is one of the most widely consumed mushrooms in the world, with over 70 countries cultivating it. China (5 million tons) is the world's leading mushroom producer, followed by Italy (762K tons) and the United States (391 tons). The majority of mushrooms are farmed in Pennsylvania in the United States.

Fungi, unlike plants, obtain their energy from decaying materials rather than sunlight, and thrive in damp settings. It is not necessary for them to be in a shaded location, although it does assist them retain moisture. When the conditions are appropriate (usually in the fall), the mycelium network produces fruiting bodies, which at first resemble pins and consist of a thin stalk and a tiny cap. The fruiting bodies soon "mushroom" despite their tiny size at first. The veil (a thin membrane beneath the cap) ruptures when the cap, which resembles an umbrella, gets large enough, allowing the gills to discharge spores. If the spores land on an adequate growth medium, they will germinate, resulting in the appearance of fungal filaments. Before fruiting, some fungi require a specific quantity of light, while others may thrive in dark caverns.

Psilocybin, a hallucinogenic chemical found in some mushrooms, has been outlawed in the United States for over 40 years. However, when different parties investigate the potential therapeutic benefits and assess them against the hazards, attitudes may shift. Even though psilocybin is prohibited in the United States, the FDA has approved a business named COMPASS Pathways' research on the use of psilocybin to treat depression. Another study is looking into the effects of psilocybin on cancer patients' anxiety and despair. After a single dose, around 80% of the cancer patients in this research experienced considerable decreases in anxiety and depression; the effects lasted for months, and there were few adverse effects. According to a 2006 study from Johns Hopkins University, study participants rated their psilocybin experience as one of the most memorable in their lives, equal to the birth of a first child or the death of a first parent. Treatment trials for alcoholism and tobacco addiction are also underway.

3. Background and Related Work:

In Indonesia, two study publications in the field of edible and toxic mushroom identification and categorization were discovered in 2017 using the new Google search engine. The Naive Bayes and Voting Feature Interval (VFI5) algorithms were utilized in the mushroom identification approach, with prediction accuracy of 99,552 percent and 84.53 percent, respectively (Bayu Mahardika Putra, 2008). (Galieh Adi, and Surya Pradana, 2016). One of the most important aspects of data mining is classification (Sang Jun Lee, and Keng Siau).

Random Forest Algorithm delivers superior results on large datasets with the same amount of characteristics, according to Jehad Ali, Rehanullah Khan, Nasir Ahmad, and Imran Maqsood (2012), but Decision Tree is the best and easiest technique for smaller datasets with fewer instances.

Beniwal, Sunita, and Das, B. (2015) used the Naive Bayes, Bayesnet, and ZeroR classification algorithms to classify mushrooms and found that the Bayesnet approach outperformed the other three classifiers.

Using diabetes datasets, nutrition datasets, ecoli protein datasets, and mushrooms datasets, A. Swarupa Rani and S. Jyothi (2016) investigated the performance of different classifier algorithms such as Naive Bayes, Multilayer perceptron Instance Based K-Nearest Neighbor (IBK), J48 Decision Tree, Simple Cart, ZeroR, CVParameter, and Filtered Classifier.

4. Methodology:

The classification process involves following steps carried throughout in this work:

1. The mushroom classification dataset was found on Kaggle, an online source.
2. Due to their insignificant contribution to mushroom classification, data pre-processing was performed, including the removal of veil type and stalk root characteristics.
3. Weka and Python were utilized to perform data analysis and categorization, as well as visualizations, for better outcomes.
4. The data was gathered in the form of cases that were correctly identified, instances that were wrongly classified, precision, recall, and the ROC curve.
5. Finally, the data is analyzed, and the optimal algorithm for the dataset is identified. Certain characteristics were also highlighted as a result of the analysis and their apparent classification results.

5. Data Analysis and Specification:

The Audubon Society Field Guide to North American Mushrooms contains descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom (1981). Each species is labeled as either definitely edible, definitely poisonous, or maybe edible but not recommended. This last category was merged with the toxic category. The data was retrieved from: https://www.kaggle.com/uciml/mushroom-classification

There were 8,124 rows and 23 columns in the data set. Each row depicted a mushroom, with labels indicating whether it was edible or deadly. There were no missing data, although many rows had a value of "?" under "stalk root." For many of the models, this column was eliminated. Furthermore, the veil type was discovered to only have one level, therefore it was deleted to reduce dimensionality. Label encoder was used to convert variable values from letters to numeric values so that the models could process the data. For each variable, a histogram was produced, and distributions were observed. The data was split into a training set and an unlabeled testing set for some algorithms.

5.1 Exploratory Data Analysis:

We used both Weka and Python to conduct our initial data analysis. During our data analysis, we attempted to identify instances based on each feature's unique value.

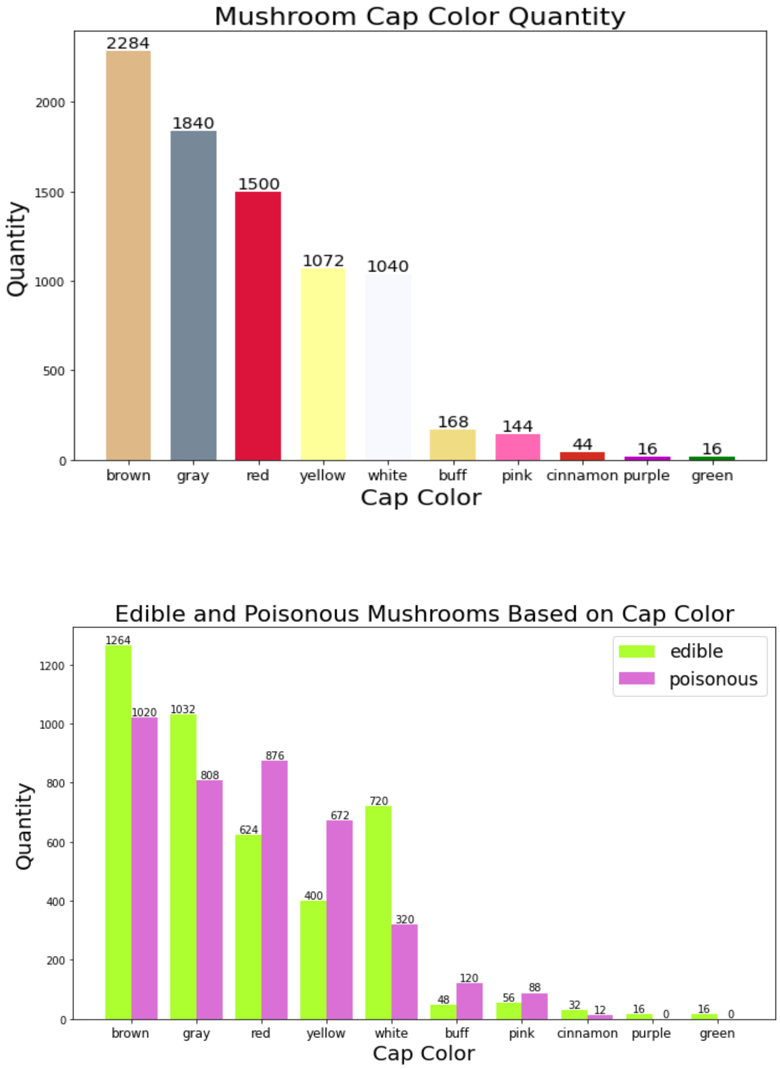

Fig 1: Instances Count Based on Cap Color & Class Distribution

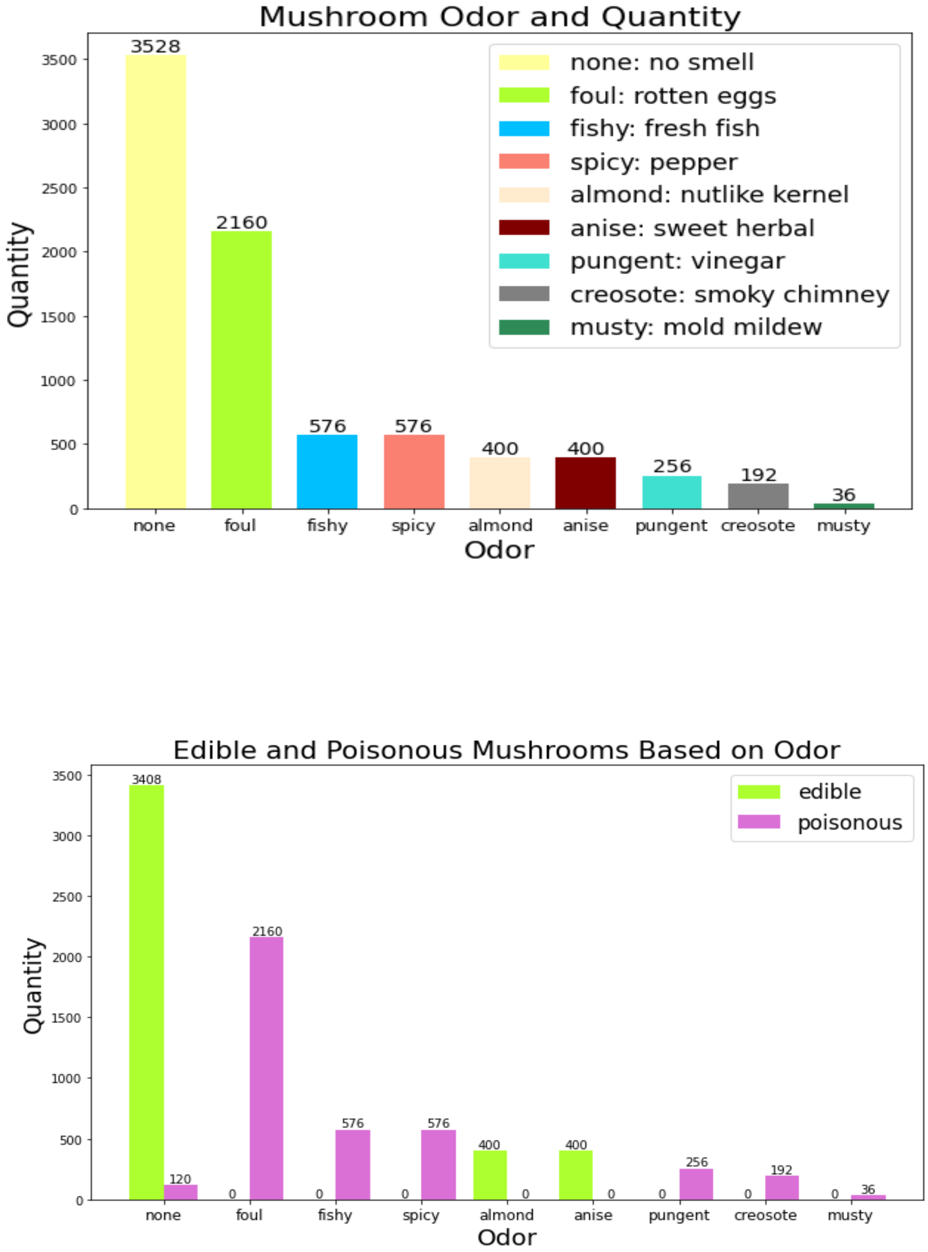

Fig 2: Instances Count Based on Odor & Class Distribution

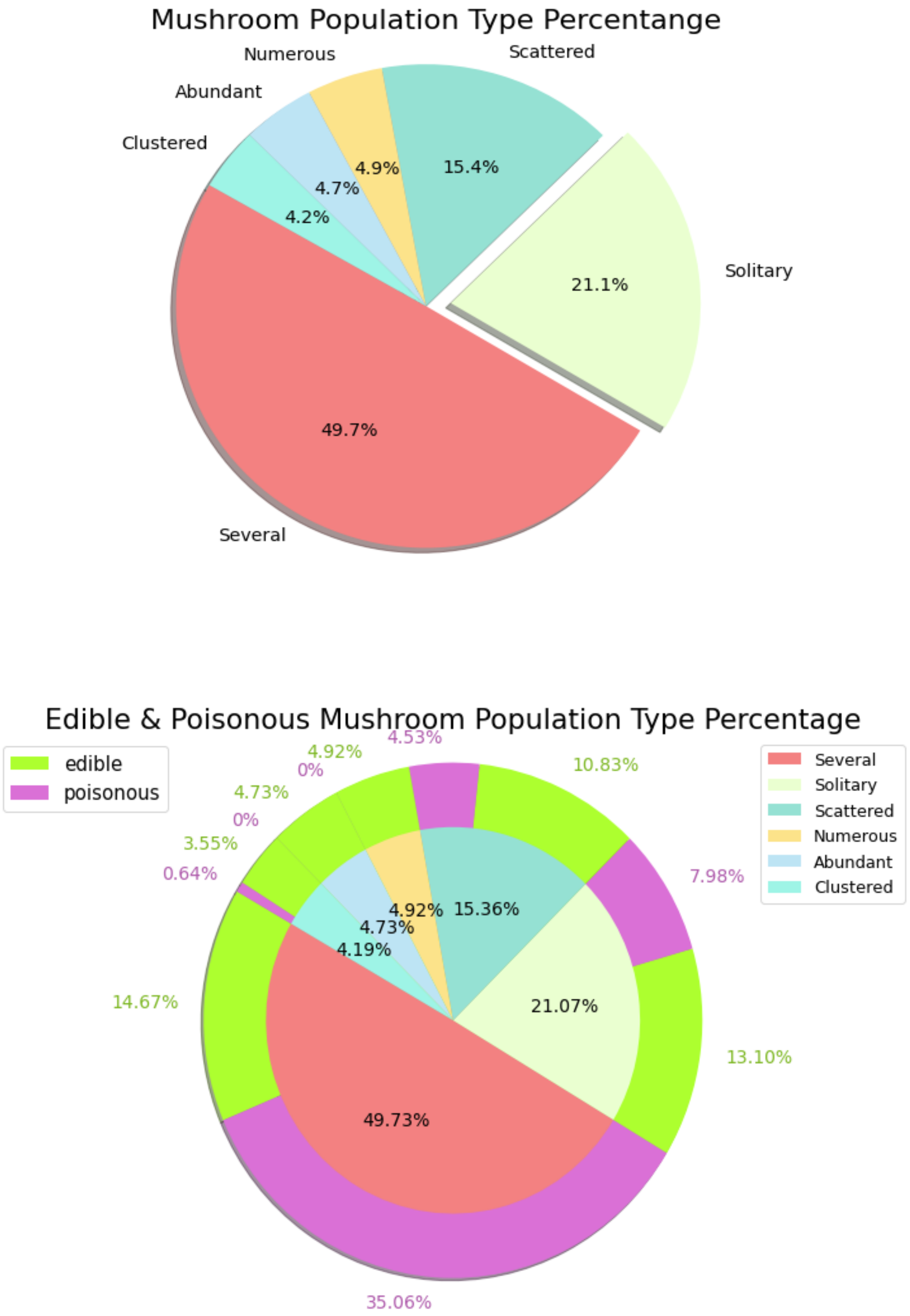

Fig 3: Instances Count Based on Population & Class Distribution

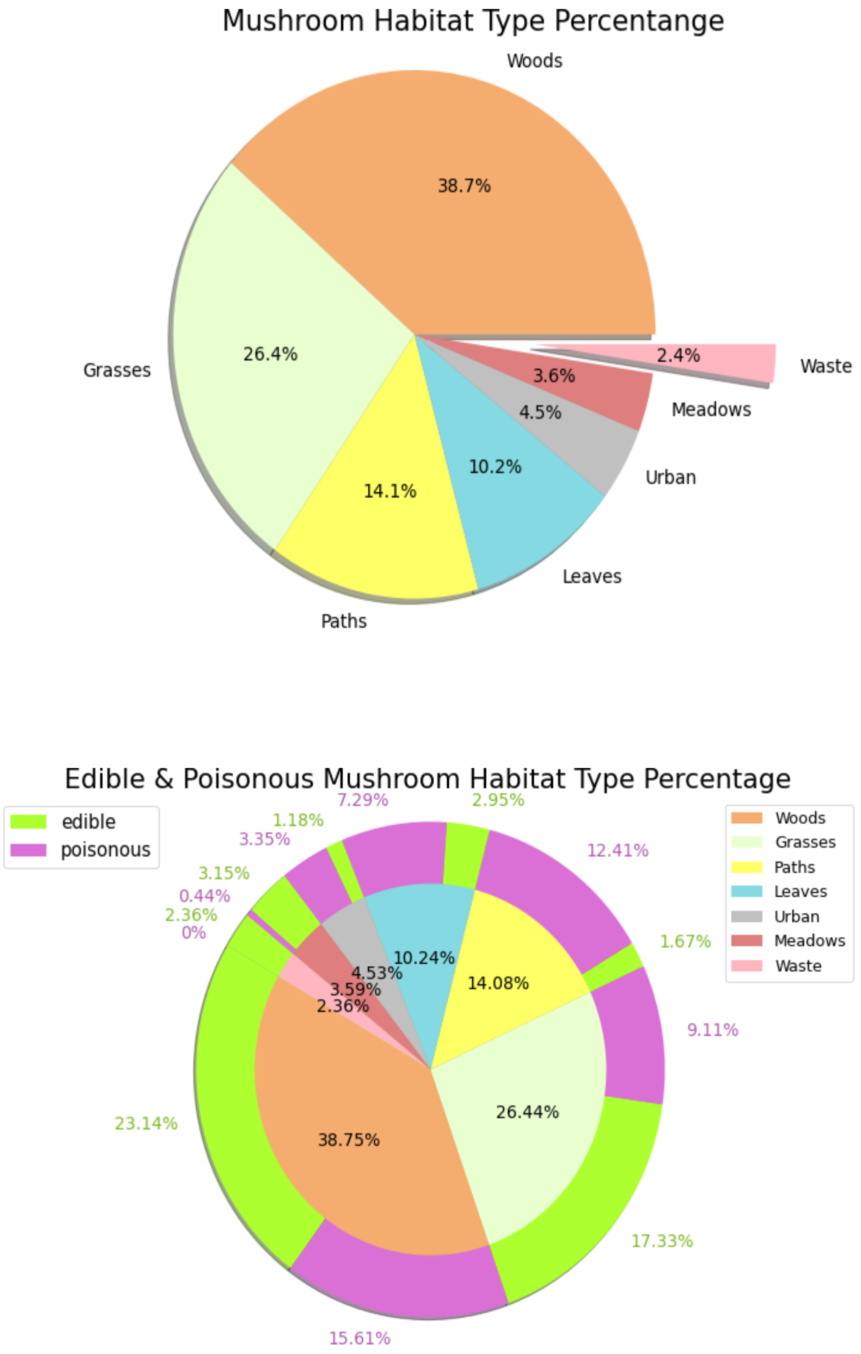

Fig 4: Instances Count Based on Habitat & Class Distribution

6. Classification Models:

6.1 C4.5 Algorithm:

The C4.5 algorithm is used to create the decision tree. This algorithm is one of the most powerful and widely used classification and prediction algorithms. The decision tree approach converts a large number of facts into a rule-based decision tree. A decision tree model is a set of rules that consider the purpose of variables to separate a large number of heterogeneous populations into smaller, more homogeneous groups. The decision tree model is more direct to probability calculation for each record on the categories or to classify records by grouping into one class since the goals of variables are often grouped definitively. Entropy and gain values are the two factors utilized to calculate the decision tree's root. Graphiz was used to create a Decision Tree graph utilizing entropy.

Entropy = $-p1 \log2 p1 - p2 \log2 p2 - \ldots$
where p1, p2... are the property of properties ranging from 0 to 1

Information Gain = IG(parent, children) = entropy(parent) − [p(c1) × entropy(c1) + p(c2) × entropy(c2) + ...]
where c1, c2... are children nodes

6.2 Naïve Bayes Algorithm:

The Naïve Bayes classification model is a statistical classification model that may be used to estimate the likelihood of belonging to a class. The Bayes theorem underpins Naive Bayes, which has classification capabilities similar to Decision Trees and neural networks. The Bayes theorem is simplified in this way by Naïve Bayes.

6.3 Support Vector Machine:

The Support Vector Machine (SVM) is a learning system that classifies data using a hypothesis room in the form of a linear function sin a high-dimensional feature space. In the SVM concept, the best separator function (hyper plane) among the limiting functions is sought. Measure the hyper plane margin and locate the maximum points to discover the optimal Hyper plane separator between the two classes. Support vector refers to the data in the divider field.

6.4 Logistic Regression:

Despite its name, logistic regression is a classification model rather than a regression model. For binary and linear classification problems, logistic regression is a simple and more efficient method. It's a classification model that's simple to implement and delivers excellent results with linearly separable classes. In the industrial world, it is a widely used categorization method. Like the Adaline and perceptron, the logistic regression model is a statistical method for binary classification that can be adapted to multiclass classification.

7. Results:

The C4.5, Logistic Regression, and SVM(SMO) algorithms show higher classification accuracy than the Naive Bayes algorithm, according to the results of testing on training data (Figure 5). Though the accuracy of C4.5, Logistic Regression, and SVM is 100 percent, C4.5 took less time than the other two models, making it a superior assessment parameter when it comes to selecting the ideal model.

For system testing or assessment, data of testing outcomes of training data is re-tested using 10-fold cross-validation. The purpose of the evaluation is to put our mining data application to the test in order to determine the percentage of system accuracy. The ten-fold cross-validation technique separates the dataset into ten pieces, nine of which will be utilized as training data and one as testing data. To acquire the average accuracy, ten training and testing processes are used. Figure 6 shows the results of the evaluation system using 10-fold cross-validation. Although the process time has increased dramatically, the accuracy of classification using the Naive Bayes algorithm has increased by 0.0615 percent, affecting the proportion of successful classification.

It was discovered that "Gill Color" was the most relevant factor in categorization using Python programming. Gill Color is a root node in the decision tree classifier; correlation analysis shows that Gill Color is the least correlated, making it the most relevant characteristic in the classification.
Another interesting result was that Decision Tree, Logistic Regression, and SVM all correctly predicted True Position and True Negative Prediction. As a result, the Precision and Recall accuracy of these classifiers, as well as the ROC curve, were at their highest levels.

Table 1: Results from Training and Train-Test Dataset

	Algorithm			
Classifier output(Training set)	**C4.5(J48)**	**Naive Bayes**	**SVM (SMO)**	**Logistic Regression**
Correctly Classified Instances	100%	95.89%	100%	100%
Incorrectly Classified Instances	0%	4%	0%	0%
Kappa statistic	1	0.9175	1	1
Mean Absolute Error	0	0.0405	0	0
Root Mean Squared	0	0.1718	0	0
Relative Absolute Error	0%	8.11%	0%	0%
Root Relative Squared Error	0%	34.37%	0%	0%
Classifier output(Train-Test split)	**C4.5(J48)**	**Naive Bayes**	**SVM (SMO)**	**Logistic Regression**
Correctly Classified Instances	100%	95.83%	100%	100%
Incorrectly Classified Instances	0%	4.17%	0%	0%
Kappa statistic	1	0.9162	1	1
Mean Absolute Error	0	0.0419	0	0
Root Mean Squared	0	0.1757	0	0
Relative Absolute Error	0%	8.40%	0%	0%
Root Relative Squared Error	0%	35.16%	0%	0%

7.1 Output from Python:

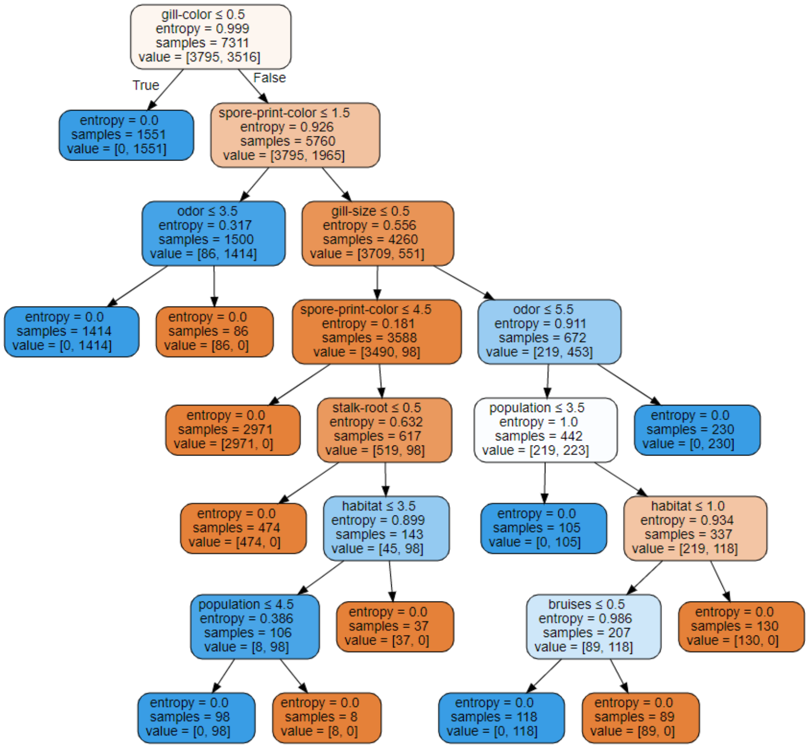

Fig 5: Decision Tree Based on Entropy: "Gill Color" Highest Entropy

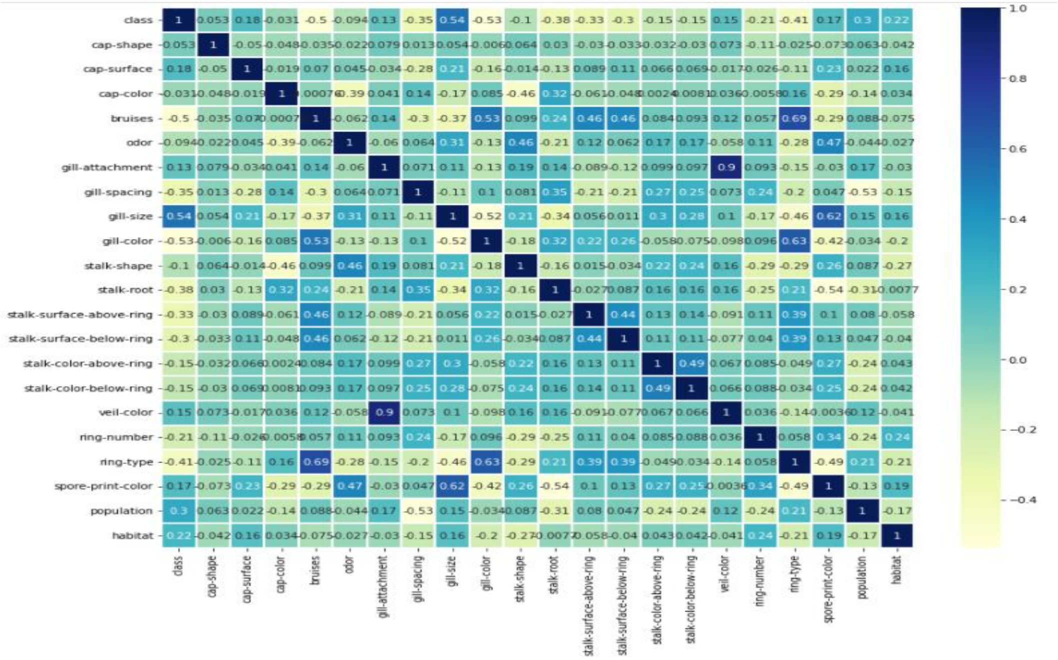

Fig 6: Correlation Graph: "Gills Color" least correlation, most important feature

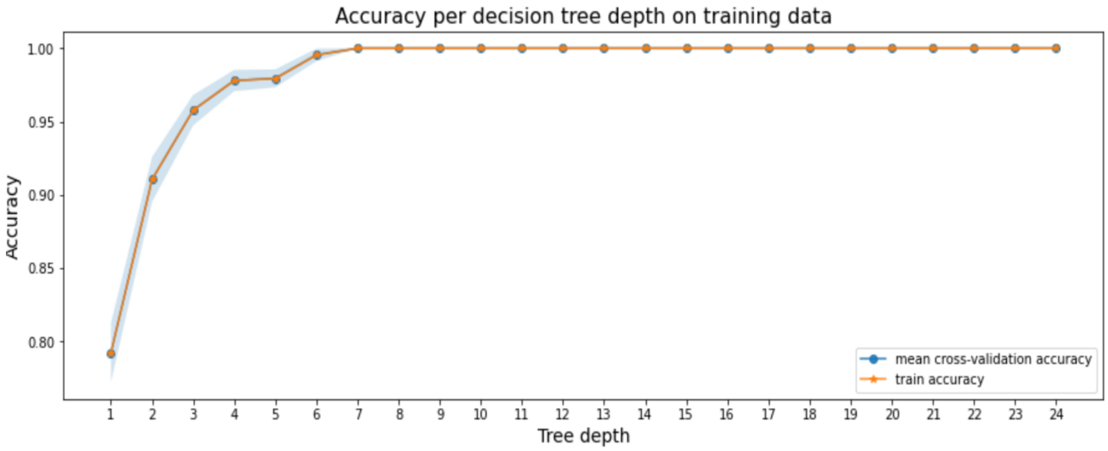

Fig 7: Tree Depth vs Accuracy(Graph tends towards perfect accuracy as depth grows)

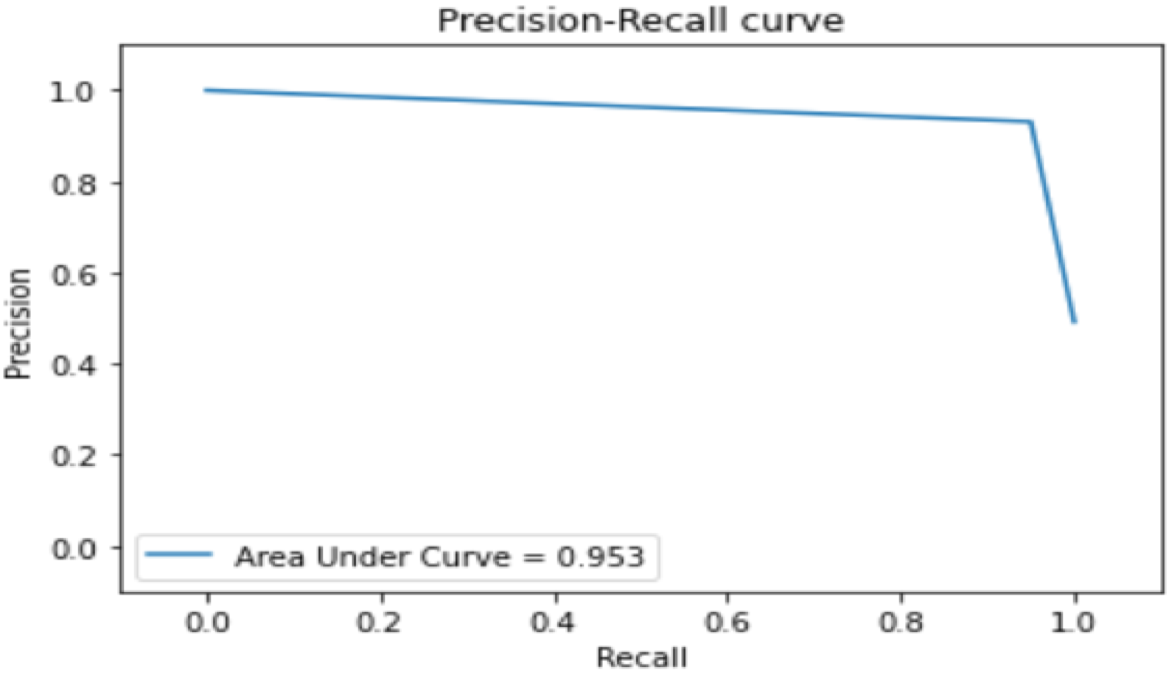

Fig 8: Precision & Recall(The rate of True Positive prediction lean towards perfect accuracy)

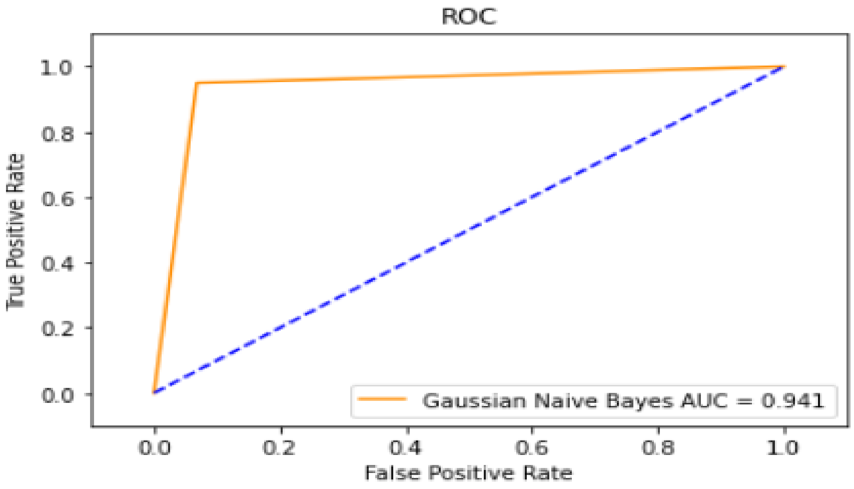

Fig 9: True Positive vs False Positive Rate (ROC curve) tends towards perfect accuracy

8. Evaluation:

It is very critical to evaluate the models that were tested on the dataset using various approaches before the model is deployed in the real world. Model should be generalized to represent the whole population as a whole. It should not be overfitting, it should be simple and easy to interpret. To evaluate models in this topic, the expected value framework is employed.

Confusion Matrix:

		Actual Values	
		P(edible)	N(poisonous)
Predicted Values	T	4208	0
	F	0	3916

↓ (Probability conversion)

		Actual Values	
		P(edible)	N(poisonous)
Predicted Values	T	0.51	0
	F	0	0.48

Cost Benefit Matrix: For cost benefit analysis, we need to understand the cost and benefits attached to mushrooms. Let's assume this data is being used by a retailer who is selling mushrooms. If the retailer is a local mushroom hunter and sells the mushrooms per pound at 2 $ and selling, therefore the profit margin would be 2$ - cost of packing which can be assumed negligible. So for every edible mushroom sold the benefit would be 0.51*2. For mushrooms that were identified poisonous, the retailer didn't stock them for sale, so the benefit would be 0$. Therefore, the cost/benefit matrix would look like:

		Actual Values	
		P(edible)	N(poisonous)
Predicted Values	Y	2$	-100$
	N	0	0

Hence, according to expected value framework which is EV = p(o1)· v(o1) + p(o2)· v(o2) + p(o3)· v(o3) + p(o4)· v(o4) = **0.51*2 -100*0 – 0*0 -0*0= 1.02**

Overall, on this dataset and the classifier with 100% accuracy the Expected Value is positive but with 1 False Negative the expected value can go down below zero drastically and can have dire consequences. Therefore, it is necessary for the classifiers to have highest accuracy possible and it should be tested with different baselines and classifiers before deploying it to the real world.

9. Conclusion:

Classification models performed really well and some of the classifiers such as Decision Tree, Logistic Regression and Support Vector Machine have perfect accuracy. Naive Bayes showed 95% accuracy which could prove deadly in case of predicated false negatives. Therefore, choosing other models that perform perfectly are best suited. In addition to being able to identify the species, several characteristics of mushrooms are clues as to whether they may be poisonous or edible. Mushrooms with no odor, or an almond or anise odor are almost always edible. Mushrooms with a foul, creosote, pungent, spicy, fishy, or musty odor are almost always poisonous. Correlation factor shows that gill color is one of the most important features in classifying mushrooms. For example: gill colors of buff or green are usually poisonous, while red or orange are usually edible.

10. References:

Mushroom Classification dataset: https://www.kaggle.com/uciml/mushroom-classification

Ramsbottom J. (1954). Mushrooms & Toadstools: A study of the activities of fungi.

Wagner, D., Heider, D. & Hattab, G(2021). Mushroom data creation, curation, and simulation to support classification tasks.
https://www.nature.com/articles/s41598-021-87602-3

Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood (2012). Random Forests and Decision Trees. IJCSI International Journal of Computer Science Issues.

Beniwal, Sunita, & Das, B(2015). Mushroom classification using data mining techniques. International Journal of Pharma and Bio Sciences.

A. Swarupa Rani and S. Jyothi (2016). Performance analysis of classification algorithms under different datasets. 3rd International Conference on Computing for Sustainable Global Development.