Mushrooms Edibility Analysis

Kyle Cheng  
500627248  
Supervisor: Tamer Abdou  
Updated for 7/17/2023

Contents

[Abstract 3](#_Toc140017633)

[Literature Review 5](#_Toc140017634)

[Topic Analysis 5](#_Toc140017635)

[Description of the Dataset 11](#_Toc140017636)

[GitHub Repository Link 13](#_Toc140017637)

[Methodology 13](#_Toc140017638)

[Evaluation of Results 13](#_Toc140017639)

[Exploratory Data Analysis 14](#_Toc140017640)

[Summary Statistics 14](#_Toc140017641)

[Data Cleaning and Preprocessing 17](#_Toc140017642)

[Feature Engineering 17](#_Toc140017643)

[Influential Predictors 18](#_Toc140017644)

[Hypothesis Testing 21](#_Toc140017645)

[Dimensionality Reduction 22](#_Toc140017646)

[Data Visualization 24](#_Toc140017647)

[Initial Results 31](#_Toc140017648)

[Test/Train Data 31](#_Toc140017649)

[Logistic Regression Classification 31](#_Toc140017650)

[Naïve Bayes Model 31](#_Toc140017651)

[Decision Tree Classification 32](#_Toc140017652)

[Random Forest Classifier 32](#_Toc140017653)

[Evaluation 33](#_Toc140017654)

[General Evaluation Notes 33](#_Toc140017655)

[Logistic Regression Classification Evaluation 34](#_Toc140017656)

[Naïve Bayes Model Evaluation 34](#_Toc140017657)

[Decision Tree Classification Evaluation 35](#_Toc140017658)

[Random Forest Classifier Evaluation 35](#_Toc140017659)

[Model Ranking and Decision 36](#_Toc140017660)

[Final Results 36](#_Toc140017661)

[Key Features 36](#_Toc140017662)

[Patterns and Combinations 37](#_Toc140017663)

[Geographic Features 38](#_Toc140017664)

[Model Applicability Features 39](#_Toc140017665)

[References 39](#_Toc140017666)

[Appendix A 41](#_Toc140017667)

[Appendix B 42](#_Toc140017668)

[Appendix C 43](#_Toc140017669)

# Abstract

A problem that has persisted since the dawn of human evolution has been the ability to discern naturally occurring resources as safe or harmful for consumption. Our ancestors have had to rely on instinct, observation and most dangerously, trial and error. The consequences of ingesting toxic or poisonous substances could be severe, ranging all the way from mild sickness to loss of life. Early humans likely passed down knowledge and observed animal behaviours in mitigating the risk of this process (Rozin, 1990). As we progress to today, this collective knowledge has been passed down through traditional wisdom and now generally, food choices should be safer. Nevertheless, mistakes and accidental poisonings occur, contributing to the ongoing challenge of discerning safe foods (Wennig et al., 2020).

Of these food sources, a particularly interesting species is the mushroom. They offer an assortment of textures and flavours for the potential gourmand however at the same time, many varieties can carry toxins and chemicals with adverse side effects (Wennig et al., 2020). While we have access to detailed databases of toxic fungi, certain scenarios will hinder that knowledge from being readily available. A person in a survival situation coming across a batch of wild mushrooms would welcome a guide that determines if the fungi are safe to eat based on the organism’s features. Therefore, correlating mushroom features to its edibility would prove as a worthwhile venture to the general public.

The theme of this task will explore classification and regression. These themes are both machine learning techniques that can be used to predict categorical or continuous target variables. Classification and regression can be used in a wide range of applications and can utilize the power of decision trees and logistic regressions to arrive at a conclusion.

The problem we wish to solve is to determine what features in mushrooms are the most important for determining the toxicity of that species and conclusively advise if the fungus is safe to ingest with a certain level of confidence. Below are some research questions to consider:

* What are the key characteristics that differentiate a toxic mushroom from a non-toxic one?
* Are there specific patterns or variations that serve as strong indicators of mushroom toxicity?
* Can machine learning or AI algorithms assist in the automated identification of toxic mushrooms?
* Are there geographic features that can be incorporated into identification protocols?

By answering these questions, we aim to acknowledge if there is a relationship between various features of a mushroom and its edibility. By creating a predictive model, we can isolate the most important characteristics in determining if a mushroom is safe to eat and draw some conclusions from the results.

The dataset to be utilized will be the “Secondary Mushroom Dataset Data Set” available from the following link: <https://mushroom.mathematik.uni-marburg.de/files/>. Entries have 20 attributes that consist of the cap, gill, stem, coloring and veil features, bruising, spore color, habitat and season. Each example is also classified in an acceptability level that is either edible or poisonous, with the latter category including unknown edibility. With over 60000 instances, the recordings provide a robust and diverse dataset that can be used to analyze what key factors contribute to a mushroom’s toxicity levels.

Some techniques and tools we are proposing to solve the stated problem include a correlation analysis, principal component analysis (PCA) and predictive models such as a decision trees or the Naïve Bayes test.

# Literature Review

## Topic Analysis

The mushroom is type of fungi that encompasses a vast array of the category. There is a huge variety of fungi estimated to range from 500,000 to 9.9 million species. Of this value, only a relatively small fraction of approximately 74,000 to 120,000 have been identified and formally named (Hawksworth, 2001). This diverse number of species highlights the inevitable challenge in creating an all-encompassing guide for mushroom identification and information. To tackle this demand, the proposed solution will focus around creating a model that identifies key characteristics of a mushroom and conclude if that organism is safe for consumption. Mushroom anatomy is informally associated by two parts, its cap and stem. However, as every species of the fungus is unique, mycologists define the mushroom with several other identifiable features. These features can include the presence and arrangement of gills, the existence of a skirt or volva, complexity of the mycelium (the fungal network, similar to the roots of a tree), the inclusion of a veil and presence of scales, among other physical traits. Moreover, mushroom knowledge extends beyond visual identification and involves understanding the habitats and ecosystems in which different species thrive. Knowing the specific regions, seasons of growth, and environmental conditions favourable for certain mushrooms can greatly assist in their identification and minimize the risk of unintentionally consuming toxic varieties. Through the analysis of some of these existing features into comprehensive models, researchers can evaluate the safety and suitability of specific mushrooms for consumption. As an important source of nutritional value, mushroom knowledge carries increased importance when venturing into discerning its edibility. Several species have evidence of preventing cancer, increasing metabolism and improving brain function (Roncero-Ramos & Delgado-Andrade, 2017). Nevertheless, it is important to highlight that mushroom identification is a complex and intricate field. Even with knowledge and experience, accurately identifying mushrooms can be challenging, and there is always a degree of risk involved. Therefore, it is crucial to exercise caution, verify information from reliable sources, and when uncertain, seek guidance from expert mycologists or experienced foragers before consuming wild mushrooms. By expanding our knowledge about mushrooms, their characteristics, habitats, and cultural significance, we can confidently navigate the realm of mushroom foraging, ensuring a safer and more gratifying journey. The proposed model will analyze the specific combination of characteristics exhibited by each mushroom and provide a level of valuable insight into its edibility or toxicity. By developing robust models and integrating insights from mycology and related fields, we can effectively enhance our ability to classify and evaluate mushrooms for both safety and enjoyment.

Existing research focusing on identifying mushroom edibility is prevalent. Some works already exist and utilize machine learning to answer several research topics important to the study of mushroom edibility. These answers include the identification of which key features are the most prominent in toxic and non-toxic mushrooms. Secondly, another focus is the labelling of specific patterns or variations that can serve as strong indicators of mushroom toxicity. Thirdly, is how can the spores, gills and other reproductive structures occur in mushrooms can aid in identifying poisonous species. Fourthly, is if machine learning or AI algorithms can assist in the automated identification of toxic mushrooms. Lastly, the topic asks if there are geographic features that can be incorporated into identification protocols.

A notable paper is the “Prediction of Whether Mushroom is Edible or Poisonous Using Back-propagation Neural Network.,” Alkronz et al. (2019) asks and answers similar questions to the prior discussed. In this report, the goal is to train a neural network model using a back-propagation algorithm. In this report, an alternative dataset to the proposed dataset is used to create the model. Analogous attributes occur under the categories of cap features, gill features, stalk/stem features, veil features, ring features, bruising, spore colour and habitat. The divergence of similarity arises when the Alkronz et al. report includes odour and mushroom population density attributes whereas the proposed does not. In replacement of the absent features, the proposed dataset includes a column for the season in which the mushroom was observed and harvested. The key difference is that the Alkronz et al. paper’s goal is the usage of Back-propagation neural networks (BPNN).

BPNN are similar to familiar models such as decision trees and regression models. They are all forms of machine learning with each method having its own set of strengths and weaknesses that we will briefly explore. BPNN can be referred to as the "backward propagation of errors," and is widely utilized algorithm in the training of artificial neural networks. It serves as a fundamental technique for modifying the network's weights and biases based on the disparity between predicted and desired outputs (Dai & Macbeth, 1997). By iteratively applying forward propagation, error calculation, and backward propagation, the neural network gradually learns to adjust its weights and biases to minimize the error between predicted and actual outputs. This process allows the network to effectively learn from the training data and make more accurate predictions over time. In comparison to the proposed methods of using decision trees and regression models, the BPNN creates several differences to take note of. BPNNs excel at capturing intricate patterns and non-linear relationships between input features and output labels. In contrast, the other techniques, may struggle to capture non-linear relationships effectively, often relying on simpler approximations. Importantly for our target audience of the general public, decision trees and regression models, offer interpretability when presenting decision rules or feature coefficients that can be easily understood. Whereas neural networks, are considered more opaque and interpreting the relationships learned in hidden layers can be challenging. An advantage that the simpler models possess is the offering of explicit measures of feature importance. This reveals influential features based on metrics such as information gain or coefficients. On the other hand, determining feature contributions in neural networks can be more complex and less straightforward. Selecting the appropriate model depends on specific problem requirements, available data, interpretability needs, computational resources, and performance objectives. It is beneficial to experiment with different approaches, assess their performance using appropriate evaluation metrics, and choose the most suitable model for a given task.

The resulting conclusion derived from the Alkronz et al. paper finds that gill size is the most relatively important feature that determines toxicity in mushrooms. Additionally, mushroom population density, the feature absent from our proposed dataset, scored relatively high over the 75th percentile on most important features. Odour however, did not perform nearly as well sitting at the bottom 33rd percentile. This will be interesting to note as we will place some level of importance on how the season attribute will perform in consideration to these missing attributes.

Another existing piece of research can be found in the International Journal of Advanced Trends in Computer Science and Engineering titled “Classification of Mushroom Fungi Using Machine Learning Techniques,” Ottom et al., (2019). This work covers the effectiveness of using a variety of machine learning algorithms in order to classify if a mushroom is edible or poisonous. Utilizing a dataset differing from the proposed, attributes cover features such as family, location, dimensions and edibility. Evaluation of each model is then scrutinized with accuracy, precision, recall and F1-score. In this instance, support vector machines (SVM) achieve the highest accuracy followed closely by the random forest method. Based on the findings presented in the paper, it is determined that certain machine learning approaches, including SVM and random forests, exhibit effectiveness in precisely categorizing mushroom fungi as either safe or harmful using their distinctive characteristics. These techniques possess the capability to support mushroom identification and facilitate the differentiation between non-toxic and poisonous varieties.

A Support Vector Machine (SVM) is a popular machine learning algorithm used for classification and regression tasks (Noble, 2006). It excels in solving binary classification problems, where data points are divided into two distinct groups. The main concept behind SVM is to find an optimal separation boundary, known as a hyperplane, in a high-dimensional space. This boundary effectively separates the data points of different classes. Support vectors, which are the nearest data points to the separation boundary, are crucial in defining this optimal hyperplane. The goal of SVM is to maximize the margin, the distance between the separation boundary and the closest support vectors, for optimal class separation. When a linear separation boundary is insufficient, SVM employs the kernel trick. This technique projects the data into a higher-dimensional space, enabling a linear boundary to separate the classes effectively. Common kernel functions, like linear, polynomial, radial basis function (RBF), and sigmoid, are employed for this purpose. SVM offers advantages, such as handling high-dimensional data, being resistant to over fitting, and capturing non-linear relationships through the kernel trick. It finds applications in diverse domains, including image classification, text categorization, and bioinformatics.

The findings from the Ottom et al. paper conclude that SVM and random forest algorithms exhibit notable accuracy in distinguishing between the two categories of mushrooms. This research underscores the potential of machine learning techniques in mushroom classification and offers valuable insights into algorithmic performance of this specific task.

The venture of this paper will attempt to replicate the effectiveness of BPNN and SVM algorithms using:

1. Logistic Regression,
2. Decision Trees,
3. Naïve Bayes Model
4. Random Forest Classifier.

By analyzing the suitability of these models, this research seeks to unveil their potential in precisely classifying mushrooms as either edible or poisonous based on their distinctive characteristics. Moreover, the study will delve into the role of decision trees and consider their capacity to effectively capture complex relationships and handle categorical attributes. Due to the use of secondary data and differing analysis models, this paper will supplement the aforementioned works and seeks to add to the ongoing conversation. Through a comprehensive comparison of the performance of each model alongside other approaches, this research intends to reveal any unique strengths or limitations of this particular methodology. The results obtained will contribute valuable insights into identifying the most appropriate machine learning techniques for mushroom classification and shed light on the practicality of leveraging each model in this specific domain.

## Description of the Dataset

The dataset, accessible through the provided link, <https://mushroom.mathematik.uni-marburg.de/files/SecondaryData/secondary_data_no_miss.csv>, offers an expanded collection of information and attributes pertaining to mushroom fungi. It serves as a complementary resource to the primary mushroom dataset, aiming to augment the comprehension and exploration of mushroom characteristics. Within this secondary dataset, a range of features is included, encompassing diverse aspects such as cap diameter, cap shape, cap surface, cap color, bruise/bleed characteristics, gill attachment, gill spacing, gill color, stem height, stem width, stem colour, ring presence, ring type, habitat and finally, the season of observation/harvest.

By delving into the nuanced intricacies of cap diameter, shape, and surface, researchers can unravel distinctive traits that set apart individual mushroom species. The incorporation of features like gill attachment, spacing, and color lends further discernment into the structural and visual attributes of mushrooms. Moreover, the inclusion of stem height, width, and root details enriches our comprehension of the overall morphology exhibited by diverse mushroom species.

The dataset also encompasses attributes encompassing the veil, ring, and spore color, which play a pivotal role in precise identification and classification endeavours. By delving into these features, researchers can delve into the presence or absence of veils, the diverse types of rings encircling the stems, and the unique hues exhibited by spores, all of which contribute significantly to species differentiation. Furthermore, the dataset captures valuable insights into the habitats favoured by specific mushroom species, unravelling their preferences for thriving in distinct environments, be it lush forests, verdant grasslands, or even bustling urban areas.

To enhance the depth of analysis, the dataset incorporates the recorded season of observation or harvest, offering researchers a valuable lens to explore potential patterns in mushroom growth and availability across different times of the year. This temporal dimension empowers the investigation of how environmental factors and climatic conditions interplay with the development and distribution of diverse mushroom species throughout the annual cycle. By considering the seasonal aspect, researchers can unravel the intricate relationships between nature's rhythms and the fascinating world of mushrooms, uncovering hidden insights into their lifecycle and ecological dynamics.

Each of these features contributes valuable insights into the various attributes of mushroom species, thereby facilitating further analysis and classification endeavours. The primary purpose of this dataset is to support researchers and practitioners in their investigations and studies concerning the manifold attributes and properties of mushroom fungi. By incorporating this supplementary dataset into their analyses, researchers can unlock enhanced insights into the intricate characteristics and classification of mushroom fungi, thereby fostering advancements in this domain.

## GitHub Repository Link

<https://github.com/k1cheng/CIND820-Big-Data-Analytics-Project-2023>

## Methodology

The following steps and graph details the methodology this report will follow:

1. The Secondary Mushroom Dataset Data Set is downloaded from the provided link
2. The data set is cleaned and appropriate pre-processing methods are applied
3. The data will is visualized and exploratory analysis created
4. Application of the data to the models and validating the accuracies
5. Predictions and recommendations made based off results

## Evaluation of Results

The assessment technique employed in research projects plays an important role by outlining the process used to assess the validity and reliability of the results. Upon gathering the results from our models, we will validate them with a confusion matrix alongside the accuracy, precision, recall and f1-scoring metrics. Additionally, we will run a cross-validation method known as k-fold to validate the reliability of our results.

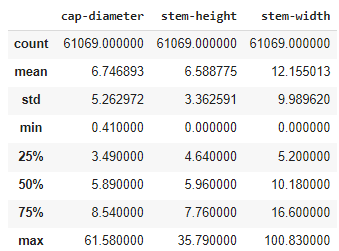
# Exploratory Data Analysis

Exploratory analysis also referred to as exploratory data analysis (EDA), will play its role in the data analysis workflow by taking the existing dataset and breaking them down into their fundamental attributes, trends, and interconnections. The principal objective of the EDA is to obtain valuable insights and devise hypotheses that can steer informed decision-making.

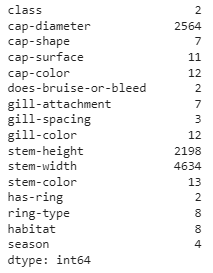
Typically, exploratory analysis encompasses a blend of statistical methodologies, data visualization techniques, and data manipulation approaches to delve into and condense the data. Code for the EDA is provided as a supplementary file in the GitHub repository link.

## Summary Statistics

This portion looks at a summary of several statistics pertaining to the dataset. First we can run a report of numerical values by using the describe() function. The below table indicates values found by this method:

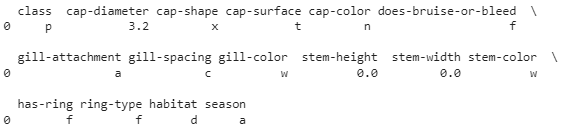


The following portion will identify the number of unique variables using the nunique() function. For example, we can see that our categorical value with the highest amount of unique values would be the stem colour:



From the above, we can determine that no column needs to be dropped as there are a minimum of at least two differing variables. If we needed to drop a column we could use the drop function again:

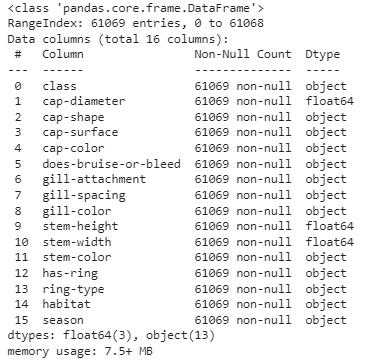
*df = df.drop([‘column\_name’],axis=1)*

Moving forward, using the mode() function allows us to see the values with most occurrences. For example, our class value carries more poisonous readings than edible: 

Addressing the above, it is important to note that certain species of mushroom can occur without a stem hence the nil readings for stem height and stem colour are not a lack of data but the lack of a physical characteristic. Otherwise, we could drop those rows with the drop() function:

*df =* *df[df[‘column\_name’]!=0*

Additional info can be found with the info() function to display the data type of each column. Moving forward, using the mode() function:



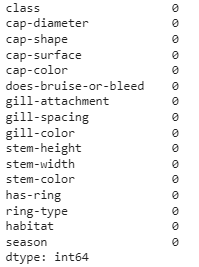
We will need to address the data type in the next section given the above findings.

Finally, we will wrap up the summary statistics section by checking if our class category is balanced using the value\_counts() function. In this case, the dataset is balanced:



## Data Cleaning and Preprocessing

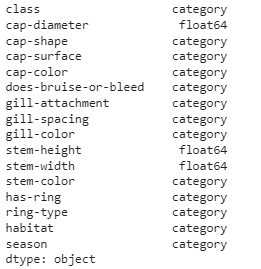
To start with the data cleaning process, we will first identify any missing variables with the isnull() function:

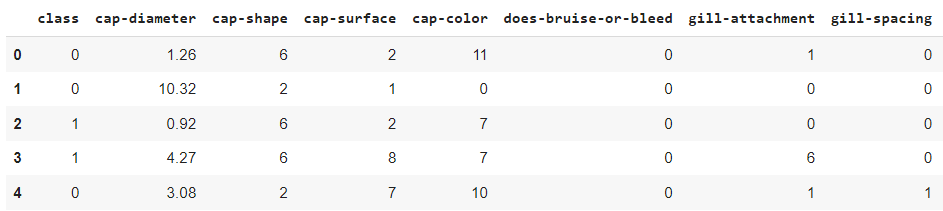


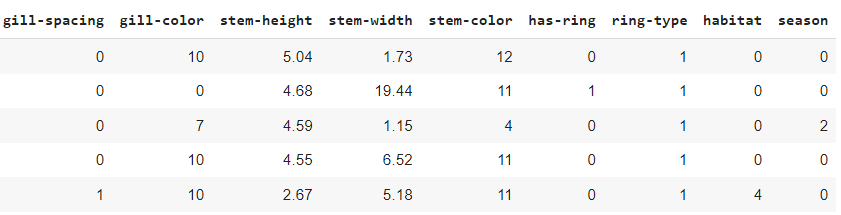
This data set has no missing values therefore no additional action needs to be taken. However, if any variables were missing, we can drop the rows that do not meet our standards with the drop() function.

## Feature Engineering

Using the info() function, we were able to ascertain that most of our values were categorical. We will convert these into an ordinal data type so that we can apply is in our models later on:

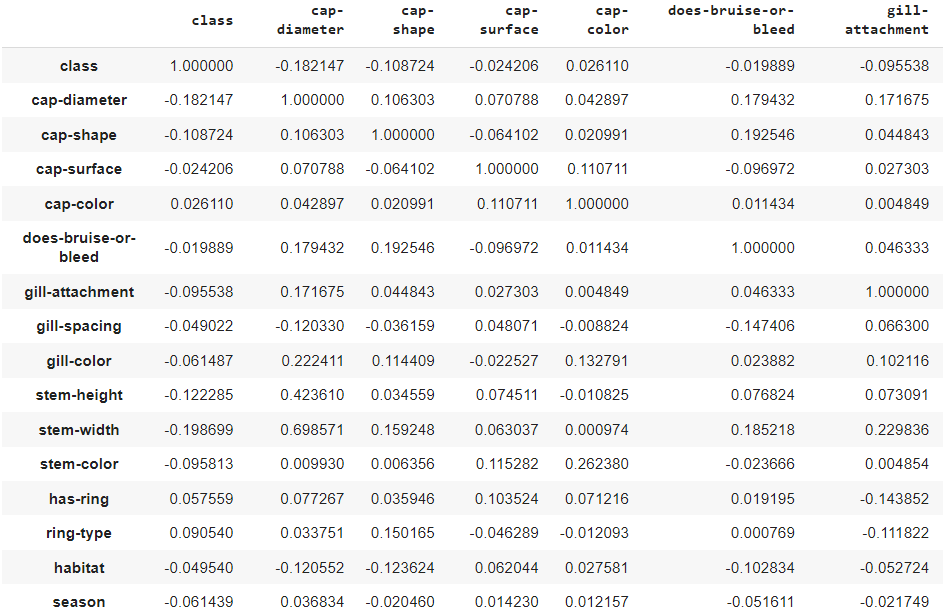


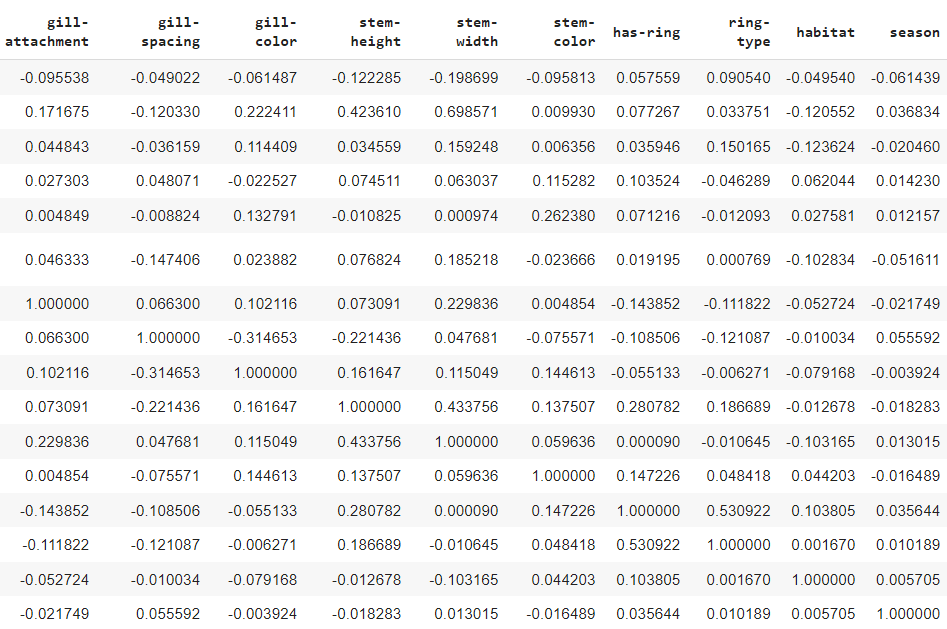




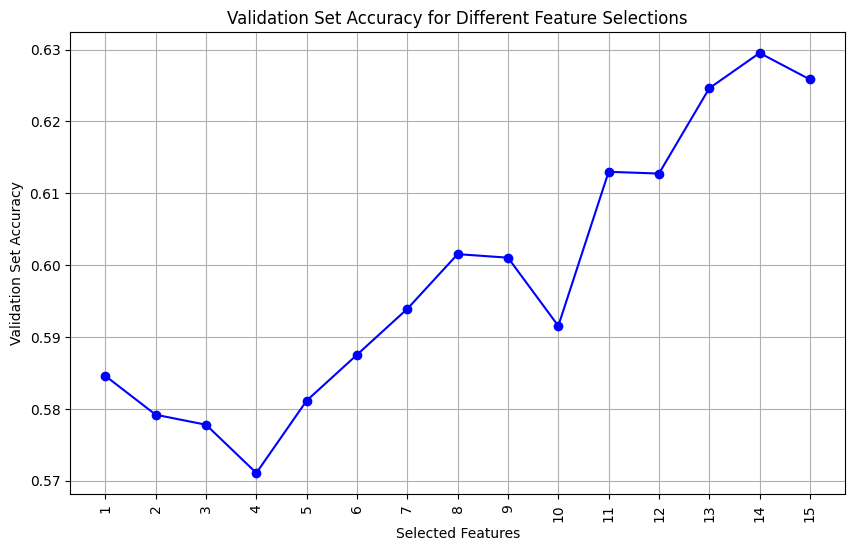
## Influential Predictors

By running a correlation matrix, we can quickly determine key features that correlate positively or negatively with our class variable. In order to do so, we will run the corr() function as below:





As we can observe, under the ‘class’ specification, all the variables are not strongly correlated in either direction. However, the highest level of correlation is found under the stem width and the cap diameter features with a correlation of -0.197 and -0.185 respectively. Again, because the variables are not strongly correlated and done so at an even rate, we will not be placing additional attention to any key features. Moving forward, we will utilize a forward selection process to better validate accuracy and influential predictors. The reported accuracy values represent the model's ability to predict the target variable, based on the chosen features. The interpretation of the results is as follows:



The provided results exhibit a progressive trend where the validation set accuracy improves as additional features are incorporated. This suggests that the additional features positively influence the model's performance. Initially, the accuracy of 0.5838 serves as the baseline when only the "cap-diameter" feature is considered. Subsequent inclusions of features like "cap-shape", "cap-surface", and "cap-color" result in marginal enhancements, maintaining proximity to the baseline. Further improvements are observed with the addition of informative features such as "does-bruise-or-bleed", "gill-attachment", "gill-spacing", and "gill-color". These additions contribute to the accuracy enhancement. As more features like "stem-height", "stem-width", "stem-color", "has-ring", "ring-type", "habitat", and "season" are integrated, the accuracy continues to rise. The highest accuracy achieved is 0.6268 when all features are incorporated.

These results demonstrate that as more features are considered, the model's accuracy fluctuates towards an increased level of accuracy. Some feature combinations resulted in improved accuracy, while others led to slight decreases. It is important to note that the specific outcomes are based on the dataset and the chosen classification algorithm. We can ascertain that the inclusion of additional features generally tends to yield more accurate results.

## Hypothesis Testing

Several of the below research questions will be elaborated upon with hypothesis testing. We will reiterate the questions from above then write out the null and alternative hypothesis as follows:

1. What are the key characteristics that differentiate a toxic mushroom from a non-toxic one?

**H0**: There are no significant differences in the characteristics found between toxic and non-toxic mushrooms.

**Ha**: There are significant differences in the characteristics found between toxic and non-toxic mushrooms.

1. Are there specific patterns or variations that serve as strong indicators of mushroom toxicity?

**H0**: here are no specific patterns or variations that serve as strong indicators of mushroom toxicity.

**Ha**: There are specific patterns or variations that serve as strong indicators of mushroom toxicity.

1. Are there geographic features that can be incorporated into identification protocols?

**H0**: There are no geographic features that can be incorporated into identification protocols.

**Ha**: There are geographic features that can be incorporated into identification protocols.

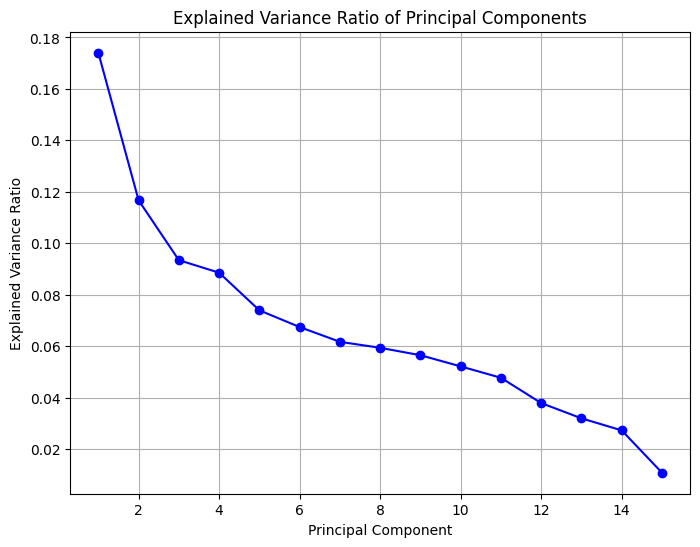
1. Can machine learning or AI algorithms assist in the automated identification of toxic mushrooms?

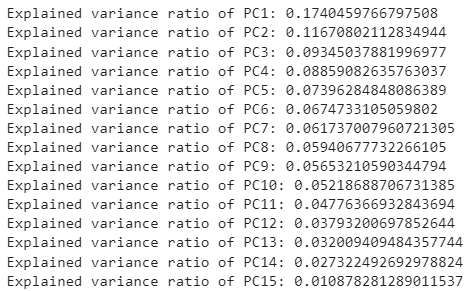
**H0**: Machine learning or AI algorithms cannot assist in the automated identification of toxic mushrooms.

**Ha**: Machine learning or AI algorithms can assist in the automated identification of toxic mushrooms.

## Dimensionality Reduction

After performing a dimensionality reduction technique called Principal Component Analysis (PCA) on the dataset, we generated a plot to visualize the explained variance ratio of the principal components. This plot provides insights into the amount of variance captured by each principal component:





The plot reveals a gradual decrease in the explained variance ratio as we move from the first principal component (PC1) to the fifteenth principal component (PC15). The initial components contribute significantly to the total variance, while the later components have diminishing contributions. The most influential component, PC1, accounts for a substantial portion of the variance at 17.40%. The second component, PC2, explains an additional 11.67% of the variance. Together, PC1 and PC2 capture approximately 29.07% of the total variance. To determine the appropriate number of principal components to retain, researchers often look for an 'elbow point' or a significant drop in the explained variance ratio. However, in this case, there is no clear elbow point, and the explained variance ratio declines gradually. Thus, the selection of principal components depends on the desired level of variance retention and the trade-off between dimensionality reduction and information loss. In summary, the PCA analysis demonstrates the potential to capture the dataset's variability using a smaller number of principal components. The plot of explained variance ratio provides valuable insights into the contribution of each component to the total variance. Moving forward, the selected principal components will be utilized for further analysis and modeling, facilitating the exploration of relationships and patterns related to mushroom toxicity prediction.

## Data Visualization

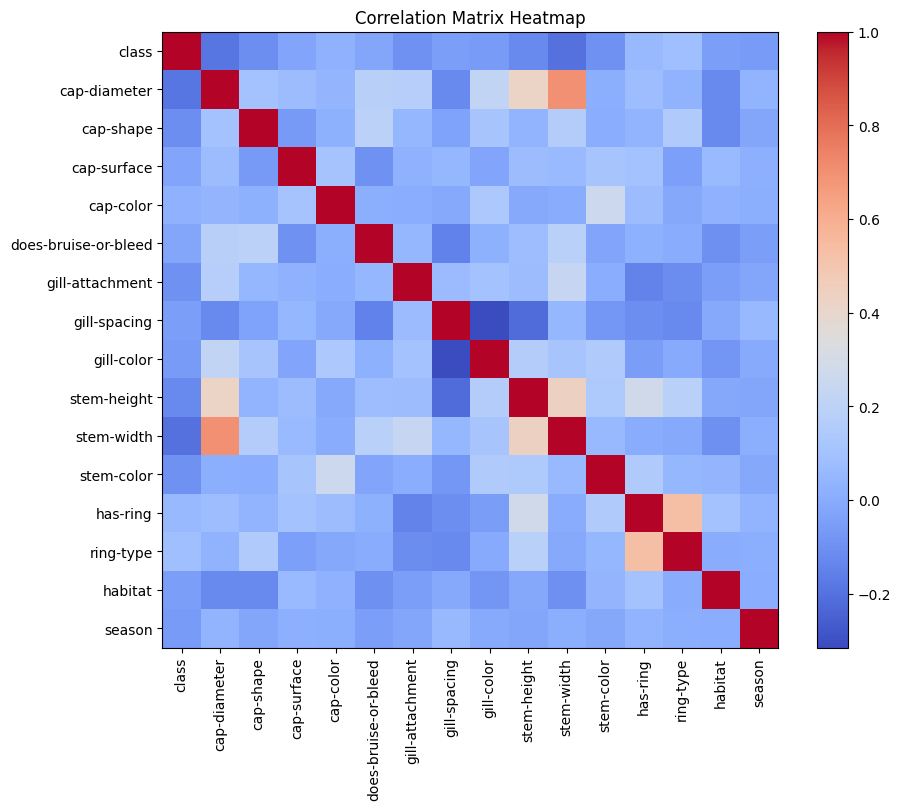
In this portion, we will utilize histograms for each feature and a correlation heat map to better represent our data outside of a list of variables.

|  |  |
| --- | --- |
| C:\Users\Yuketo\Desktop\download (4).png | * The most frequently occuring cap diameter measurement is around 3cm * The least frequently occuring cap diameters are over 25cm |
| C:\Users\Yuketo\Desktop\download (5).png | * The most frequently occuring cap shape is convex * The least frequently occuring cap shape is conical |
| C:\Users\Yuketo\Desktop\download (6).png | * The most frequently occuring cap surface is sticky * The least frequently occuring cap surface is leathery |

|  |  |
| --- | --- |
| C:\Users\Yuketo\Desktop\download (7).png | * The most frequently occuring cap colour is brown * The least frequently occuring cap colour is blue |
| C:\Users\Yuketo\Desktop\download (8).png | * The most frequent type of mushroom does not bruise or bleed * The least frequent type of mushroom does bruise or bleed |
| C:\Users\Yuketo\Desktop\download (10).png | * The most frequently occuring gill attachment is adanate * The least frequently occuring gill attachment is none |
| C:\Users\Yuketo\Desktop\download (11).png | * The most frequently occuring gill spacing is close * The least frequently occuring gill spacing is none |
| C:\Users\Yuketo\Desktop\download (12).png | * The most frequently occuring gill colour is white * The least frequently occuring gill colour is buff or brownish yellow |
| C:\Users\Yuketo\Desktop\download (13).png | * The most frequently occuring stem height measurement is around 4 to 7cm * The least frequently occuring stem heights are over 20cm |

|  |  |
| --- | --- |
|  | * The most frequently occuring stem width measurements are under 10 cm * The least frequently occuring stem widths are over 60cm |
| C:\Users\Yuketo\Desktop\download (15).png | * The most frequently occuring stem colour is white * The least frequently occuring stem colour is buff or brownish yellow |
| C:\Users\Yuketo\Desktop\download (16).png | * The most frequently occuring ring presence is false * The least frequently occuring ring presence is true |

|  |  |
| --- | --- |
| C:\Users\Yuketo\Desktop\download (17).png | * The most frequently occuring ring type is none * The least frequently occuring ring type is movable |
| C:\Users\Yuketo\Desktop\download (18).png | * The most frequently occuring habitat is in the woods * The least frequently occuring habitat is an urban environment |
| C:\Users\Yuketo\Desktop\download (19).png | * The most frequently occuring season of collection is autumn * The least frequently occuring season of collection is spring |

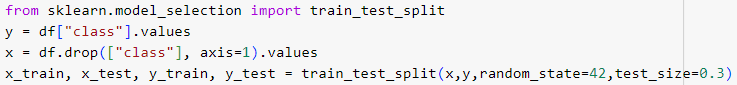


The correlation heat map above indicates the strongest positive correlation between cap diameter and stem width. Additionally, the strongest negative correlation exists between gill colour and gill spacing. When looking specifically at the class variable and other categories, the correlation could help us predetermine important features. However, the heat map shows no strong correlations in this area therefore more investigation would be required to answer our questions.

# Initial Results

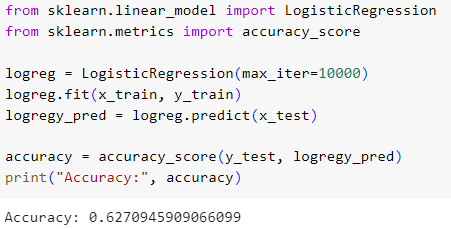
## Test/Train Data

In preparation for model selection, we will utilize the train\_test\_split method to create our training and testing data. As a baseline, we have used a 0.3 or 30% sample size for our tests. The testing and training data will include the same parameters for each modeling method.



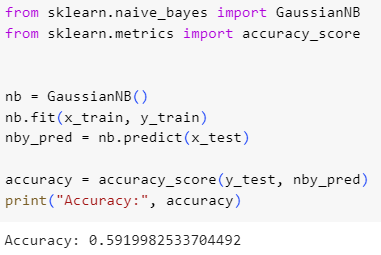
## Logistic Regression Classification

In this section, we utilized the logistic regression method and tested the accuracy:



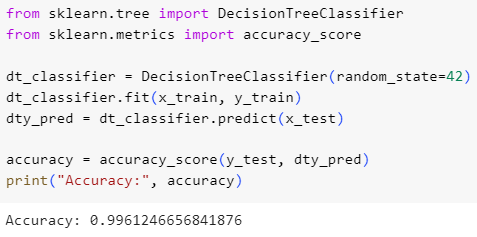
## Naïve Bayes Model

In this section, we utilized the Naïve Bayes method and tested the accuracy:



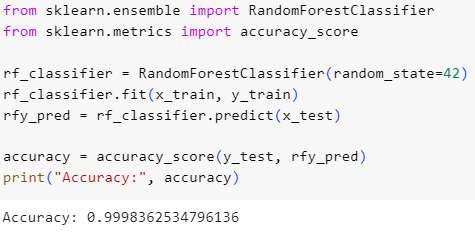
## Decision Tree Classification

In this section, we utilized the decision tree method and tested the accuracy:



## Random Forest Classifier

In this section, we utilized the random forest method and tested the accuracy:



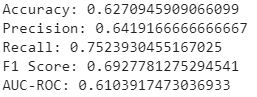
# Evaluation

## General Evaluation Notes

While accuracy, the straight forward overall correctness of a model, can be easily defined, utilizing more metric testing will offer us a bigger picture. Namely, precision, recall, F1-score and AUC-ROC will be our primary means of evaluation. Precision measures the proportion of true positive predictions among the total positive predictions made by the model. It focuses on the accuracy of positive predictions. Precision is calculated by dividing the number of true positives by the sum of true positives and false positives. Recall evaluates the model's ability to correctly identify positive instances. It represents the proportion of true positive predictions among all actual positive instances. Recall is computed by dividing the number of true positives by the sum of true positives and false negatives. The F1 score combines precision and recall into a single metric to provide a balanced measure of the model's performance. It is calculated as the harmonic mean of precision and recall, giving equal weight to both metrics. The formula for F1 score is 2 \* (precision \* recall) / (precision + recall). AUC-ROC (Area under the Receiver Operating Characteristic Curve) measures the model's ability to distinguish between classes by plotting the True Positive Rate (Recall) against the False Positive Rate. It represents the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance. AUC-ROC provides a comprehensive assessment of the model's performance across various classification thresholds.

## Logistic Regression Classification Evaluation

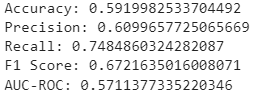
Upon initial readings, the accuracy of the logistic regression method was low at 62.71%. In order to better understand the model, we looked at the confusion matrix testing that included the accuracy, precision, recall, F1-score and AUC-ROC. The following results are below:



From the results above, we can discern the model’s validity with the F1-score of 69.28% which is not extremely low but lacking more desirable results

## Naïve Bayes Model Evaluation

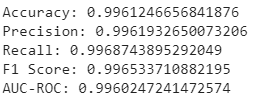
Upon initial readings, the accuracy of the Naïve Bayes method was low at 59.20%. In order to better understand the model, we looked at the confusion matrix testing that included the accuracy, precision, recall, F1 score and AUC-ROC. The following results are below:



From the results above, we can discern the model’s validity with the F1-score of 67.22% which is lower than the logistic regression classification model.

## Decision Tree Classification Evaluation

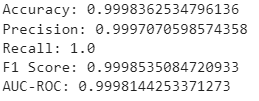
Upon initial readings, the accuracy of the decision tree method was high at 99.61%. In order to better understand the model, we looked at the confusion matrix testing that included the accuracy, precision, recall, F1 score and AUC-ROC. The following results are below:



From the results above, we can discern the model’s validity with the F1-score of 99.65% which is very high.

## Random Forest Classifier Evaluation

Upon initial readings, the accuracy of the random forest method was high at 99.98%. In order to better understand the model, we looked at the confusion matrix testing that included the accuracy, precision, recall, F1 score and AUC-ROC. The following results are below:



From the results above, we can discern the model’s validity with the F1-score of 99.98% which is extremely favourable and higher than the decision tree classification method.

## Model Ranking and Decision

With the random forest classifier in the clear lead in all metrics, we will move forward in answering our questions using this method. Below are the rankings when considering all metrics:

1. Random Forest Classifier
2. Decision Tree Classification
3. Logistic Regression Classification
4. Naïve Bayes Model

# Final Results

## Key Features

By utilizing the random forest classifier, we determined there are key characteristics that differentiate a toxic mushroom from a non-toxic one. The following list represents a feature’s level of importance in descending order:

stem-height: 0.14461048274262384

cap-shape: 0.09375286337359055

gill-color: 0.0931666003096277

stem-width: 0.08621129050950947

gill-spacing: 0.0857581742937454

does-bruise-or-bleed: 0.08555877947749156

cap-surface: 0.06088281845765875

cap-diameter: 0.05913975764133062

gill-attachment: 0.05024421674573872

has-ring: 0.0480585368006724

cap-color: 0.03520634643354958

ring-type: 0.03351114892828478

stem-color: 0.026130500969542034

habitat: 0.01495184714058319

With these findings, we can conclude that there are significant differences in the characteristics found between toxic and non-toxic mushrooms. Mainly, according to the random forest classifier, the stem height, cap shape and gill colour, in that order, carry the most significance in determining if a mushroom is poisonous or edible.

## Patterns and Combinations

In determining if there are specific combinations of features that indicate mushroom toxicity, we will apply an iterative model in determining which of these features are most important. As we have 15 features, the code to run the full amount of combinations takes unreasonably long to process. For this instance, we have selected a max number of variables of to up to 3 features.

*Note: The corresponding codes in the .ipynb file still take an extended amount of time to process. For reader convenience, we have run the codes and pasted them at the end of this document. Appendix A shows the initial code to run all combinations+ several results while appendix B shows the code that picks the top three in accuracy score.*

Below are the top three accurate combinations of features that help determine if a mushroom is toxic or edible:   
**Rank: 1**

Features: ('cap-surface', 'stem-height', 'stem-width')

Accuracy: 0.8741880901697505

**Rank: 2**

Features: ('cap-surface', 'gill-color', 'stem-width')

Accuracy: 0.862889580263086

**Rank: 3**

Features: ('cap-surface', 'gill-attachment', 'stem-width')

Accuracy: 0.855793897713007

With these findings, we can conclude that when capped at a combination of three variables, there are specific patterns of variations that can serve as strong indicators of mushroom toxicity. Cap surface, stem height and stem width together generate an 87.42% accuracy in determining the class. This method however, does not account for the maximum amount of combinations and does not isolate exactly what variable per feature accounts for the class type. Further testing is possible to obtain more comprehensive results but will require exponentially more time to run the full possibilities of combinations. Not to mention the full list of possibilities of variable combinations from each feature on top of that.

## Geographic Features

As calculated earlier, habitat did not end up being a key feature in determining the class variable. With a score of 0.01, there was not much this feature influenced. The setback of answering this question with the available data and testing is that it remains inconclusive. While we have a good set of physical features, we lack additional geographical features such as elevation, proximity to specific wild life, proximity to bodies of water, humidity etc. However based on this set of data, we could conclude (with a large asterisk) that there are no geographic features that can be incorporated into identification protocols.

## Model Applicability Features

When looking at the validity of the models we can take our readings and conclude if machine learning or AI algorithms can assist in the automated identification of toxic mushrooms. As a refresher, the model F1 scores were 0.6927, 0.6722, 0.9965, 0.9998 on the logistic regression model, Naïve Bayes model, decision tree and random forest model respectively. The latter two showed the most promising results. Therefore, after validating the results, we can conclude machine learning’s ability to assist in the identification of toxic mushrooms. The setback of this conclusion is that the stability of this model was not conducted. Appendix C shows an example of a k-fold cross validation method these models could have included however the time this took to process was similarly extended as the combination of features portion and only a k value of 2 was selected. The further increase in this value would result in a further stability of the model otherwise.

# References

Alkronz, E. S., Moghayer, K. A., Meimeh, M., Gazzaz, M., Abu-Nasser, B. S., & Abu-Nasser, S.

S. (2019, February). *Prediction of whether mushroom is edible or poisonous using back-propagation Neural Network*. International Journal of Academic and Applied Research. http://dstore.alazhar.edu.ps/xmlui/bitstream/handle/123456789/126/ALKPOWv1.pdf?sequence=1

Dai, H., & Macbeth, C. (1998, August 18). *Effects of learning parameters on learning procedure*

*and performance of a BPNN*. Neural Networks. https://www.sciencedirect.com/science/

article/abs/pii/S0893608097000142

Hawksworth, D. L. (2001). *Mushrooms: The extent of the unexplored potential*. International

Journal of Medicinal Mushrooms. https://www.dl.begellhouse.com/journals/708ae68d6

4b17c52,4cf0547d63409c36,503a72d82b669af6.html

Noble, W. S. (2006, December 1). *What is a support vector machine*?. Nature News.

https://www.nature.com/articles/nbt1206-1565

Ottom, M. A., Alawad, N. A., & Nahar, K. M. O. (2019). *Classification of mushroom fungi*

*using machine learning techniques*. International Journal of Advanced Trends in

Computer Science and Engineering. https://www.warse.org/IJATCSE/static/pdf/file/ijatc

se78852019.pdf

Roncero-Ramos, I., & Delgado-Andrade, C. (2017, May 1). *The beneficial role of edible*

*mushrooms in human health*. Current Opinion in Food Science. https://www.science

direct.com/science/article/abs/pii/S2214799317300632

Rozin, P. (1990, February). *Acquisition of Stable Food Prefer*ences. Nutrition Reviews.

https://bpb-us-w2.wpmucdn.com/web.sas.upenn.edu/dist/7/206/files/2016/09/117

AcqStablePrefsNR90-1qr3veq.pdf

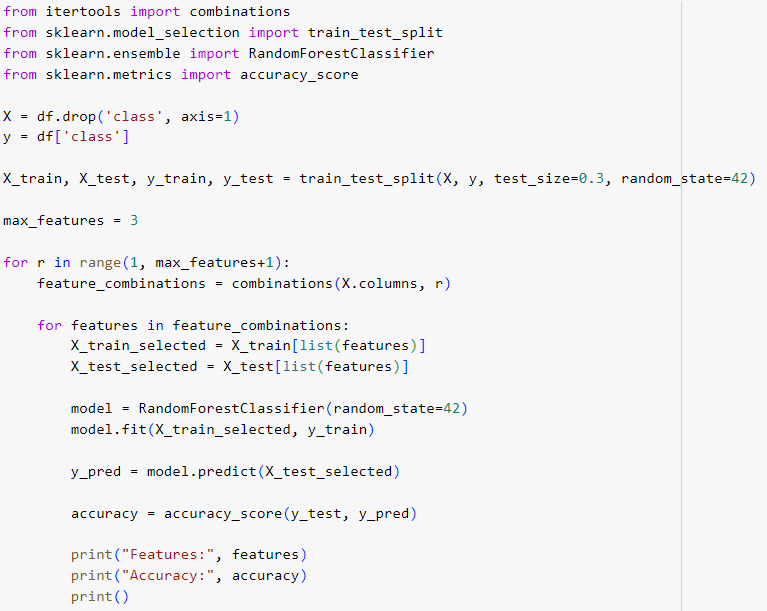
Wennig, R., Eyer, F., Schaper, A., Zilker, T., & Andresen-Streichert, H. (2020, October 16).

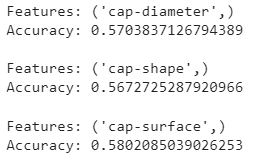
*Mushroom poisoning*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/

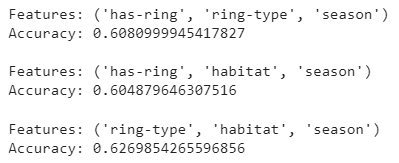
articles/PMC7868946/

# Appendix A

Iterative process showing the best combination of features and its results:

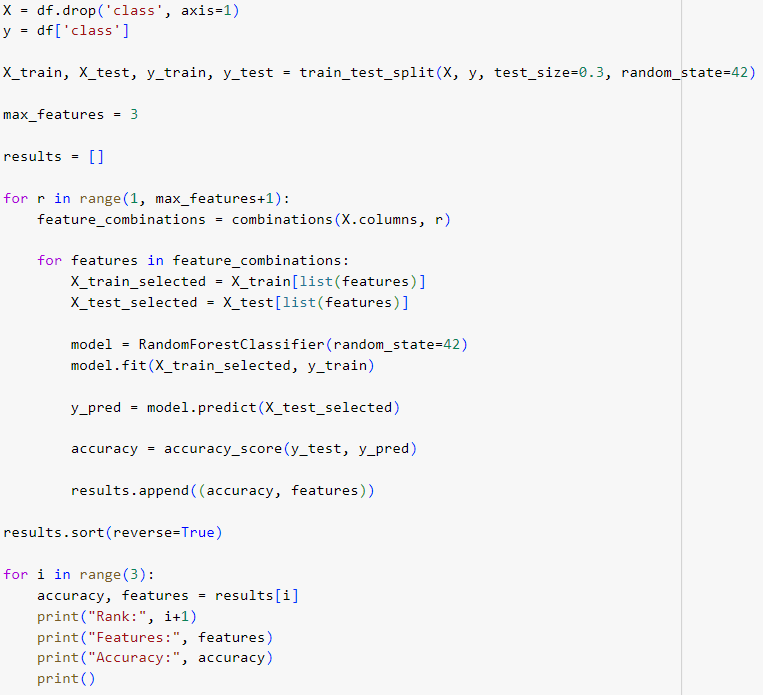


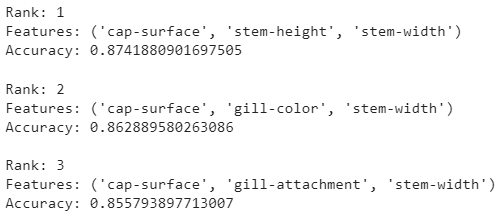




# Appendix B

Iterative process showing the best three combinations of features:





# Appendix C

k-fold cross validation for model stability:

