s**Predicting whether mushrooms are poisonous or not**

Yash Vadapalli

6507591

30th November, 2023

**Executive Summary:**

**Objective:** My main goal in this project was to create a machine learning model that can tell if a mushroom is safe to eat or not. This is really important for keeping people safe, as there are so many different kinds of mushrooms, and picking the wrong one can be dangerous.

**Methods:**

* Data Collection and Preprocessing: I used a dataset called Mushroom Records, which has lots of details about mushrooms. I got the data ready for my project by fixing any missing parts, changing categories into numbers, and making sure everything was consistent.
* Exploratory Data Analysis (EDA): I looked closely at the data to understand how different parts of it relate to each other. I put similar things together to make it easier to see patterns and figure out what's important for knowing if a mushroom is edible.
* Model Development and Training: I tried out several different ways to build a model that can learn from the data. I paid a lot of attention to how well each model worked, especially how accurate and reliable they were.
* Model Evaluation: I compared the models to see which one was best at telling if a mushroom is poisonous or not.

**Key Findings:**

* When I looked at the data, I found some really interesting patterns that helped me decide how to build the models.
* Some of the models did a better job than others. One model, in particular, was really good at predicting correctly.
* I also figured out which details about the mushrooms are most important for deciding if they're safe to eat.

**Conclusions and Recommendations:**

* I was able to make a model that's pretty good at figuring out if a mushroom is poisonous.
* I think the model could be even better if I use more data or try some more advanced techniques.
* The things I learned from this project could be really useful for teaching people about mushrooms and helping them stay safe when picking them.

This project showed me how machine learning can help us understand nature better and keep people safe. It was really exciting to see how my work could make a difference.

**Introduction:**

The primary aim of this project was to develop a machine learning model capable of accurately predicting the edibility of mushrooms. This endeavor is crucial in public health and safety, given the vast diversity and potential risks associated with various mushroom species.

Mushrooms, while a common and often nutritious addition to our diets, pose a unique challenge due to their variability. The ability to distinguish between edible and poisonous varieties is not only a matter of culinary interest but also of urgent health concern. Incorrect identification can lead to severe health consequences, making it a significant public safety issue.

In this project, the focus was on leveraging data-driven techniques to address this challenge. By creating a model that learns from existing data about mushrooms, the goal was to provide a reliable tool to assist in making informed decisions about mushroom consumption. This is particularly relevant for individuals who forage mushrooms or are faced with making choices about unfamiliar varieties.

Moreover, the project also aimed to broaden the understanding and appreciation of edible mushrooms. By providing clear insights into mushroom varieties, it hoped to reduce unnecessary apprehension around their consumption and encourage the safe exploration of their culinary potential.

Overall, this initiative represents an intersection of technology, biology, and public welfare, demonstrating the practical application of machine learning in everyday health and safety scenarios.

**Objectives:**

The project's objectives were carefully crafted to create a comprehensive machine learning model for predicting mushroom edibility. Each stage was essential in developing a reliable and informative tool:

* Comprehensive Exploratory Analysis: Initially, the project involved an in-depth analysis of the mushroom dataset to understand its characteristics and uncover patterns, laying the groundwork for model development.
* Data Preprocessing for Machine Learning: This phase focused on preparing the data by cleaning and transforming it, ensuring it was ready for effective machine learning.
* Development and Training of Classification Models: The project then moved on to build and fine-tune models that could classify mushrooms as edible or poisonous, with a focus on iterative improvement.
* Evaluation of Models: The models were evaluated on their accuracy and reliability, key factors in determining their real-world applicability.
* Extraction and Interpretation of Insights: The final objective was to derive meaningful insights from the models, enhancing understanding of what influences mushroom edibility.
* These objectives were integral to developing a tool that not only predicts mushroom edibility but also contributes to public safety and knowledge.

**Data Collection:**

The project utilized the Mushroom Records dataset, known for its detailed coverage of mushroom species. This dataset included key features like cap shape, color, gill size, and habitat, crucial for distinguishing between edible and poisonous mushrooms.

Comprising thousands of entries, the dataset was well-structured for analysis, providing a solid basis for both the exploratory analysis and the development of the predictive model. Its comprehensive nature was essential to the project's focus on accurately classifying mushrooms.

**Initial Data Exploration:**

**Loading the Dataset:**

The dataset is loaded into a pandas DataFrame named mushroom\_df.

*mushroom\_df = pd.read\_csv('mushroom.csv')*

*mushroom\_df.head()*

This code reads the mushroom dataset from a CSV file into a DataFrame and then displays the first few rows of the DataFrame using the .head() method. This is typically done to get a quick glimpse of the data format, including column names and some of the values.

**Understanding the Dataset's Size:**

Examining the shape of the DataFrame to understand the dataset's size.

*mushroom\_df.shape*

This code provides the dimensions of the DataFrame, indicating the number of rows and columns in the dataset. It's a quick way to see how large the dataset is.

**Displaying Basic Information:**

* The next step likely involves using methods like .info() or .describe() to display basic information about the DataFrame.
* While the exact code isn't shown in the snippet, such methods are commonly used to understand the data types of each column, count of non-null values, and basic statistical summaries for numerical columns.

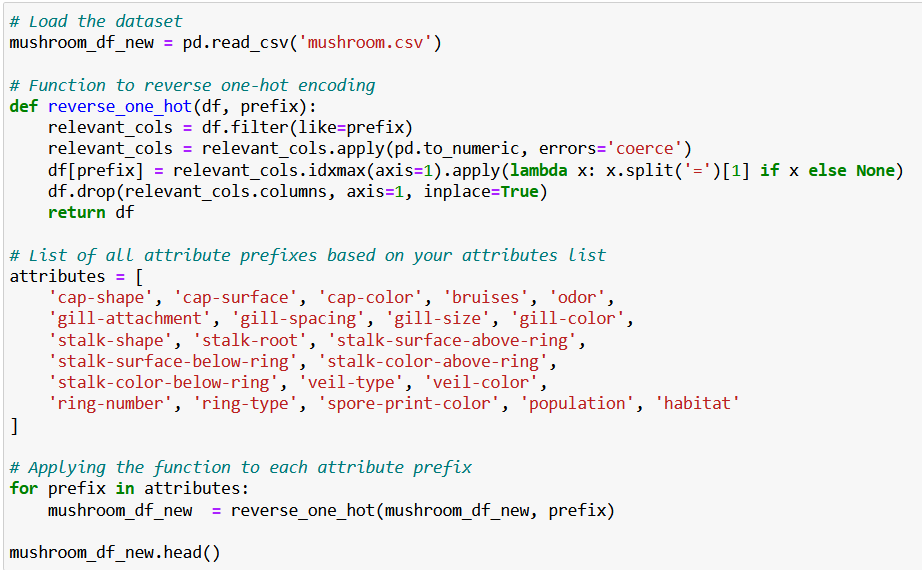
These steps mark the beginning of exploring the dataset. They are essential for getting a preliminary understanding of the data's structure, size, and the type of information each column holds. The initial exploration sets the stage for more detailed analysis, which might include checking for missing values, understanding the distribution of data, and preparing the data for further steps like data cleaning or feature engineering.

**Exploratory Data Analysis:**

In the exploratory data analysis phase, several steps were undertaken to understand the dataset better. Here's an overview of the process and insights gathered, along with suggestions on where to insert graphs:

* **Reversing One-Hot Encoding:**

To make the EDA more intuitive, the dataset was transformed back to its original categorical format from one-hot encoding.



A screenshot of a computer

Description automatically generated

This allowed for a clearer visualization of the distribution of each feature.

**Summary of Categorical Features:**

A summary of each categorical feature, including the class (edible or poisonous), was generated.

A white background with black text

Description automatically generated

This step helped in understanding the frequency of each category within the features.

**Visualizing Feature Distributions:**

Graphs were created to visualize the distribution of each categorical feature.

A screenshot of a computer code

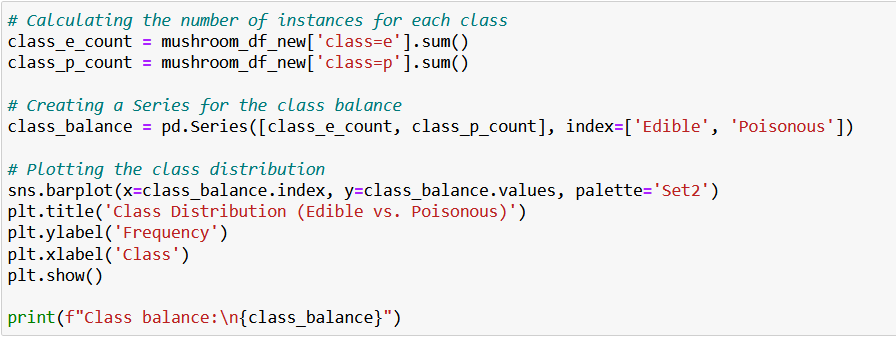
Description automatically generated

A group of graphs showing different colored squares

Description automatically generated with medium confidence

**Class Distribution Analysis:**

The balance between the two classes (edible and poisonous) was examined.

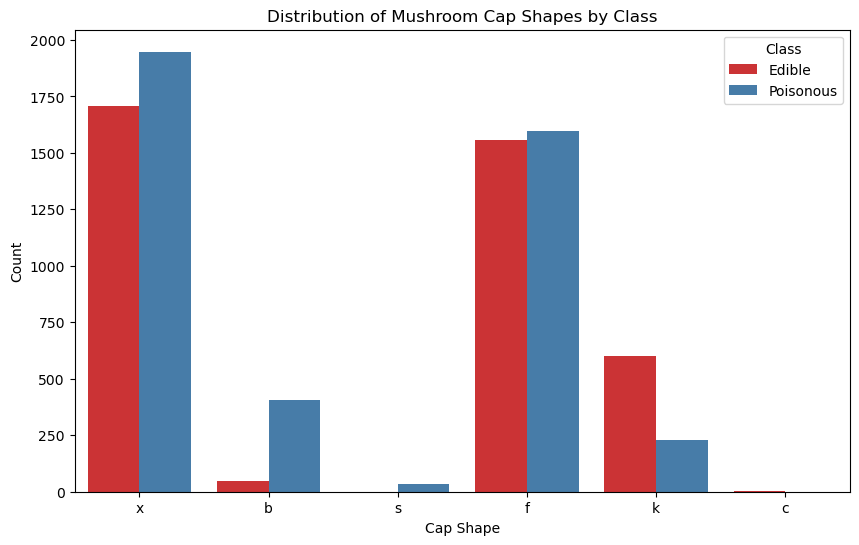


A graph of a class distribution

Description automatically generated

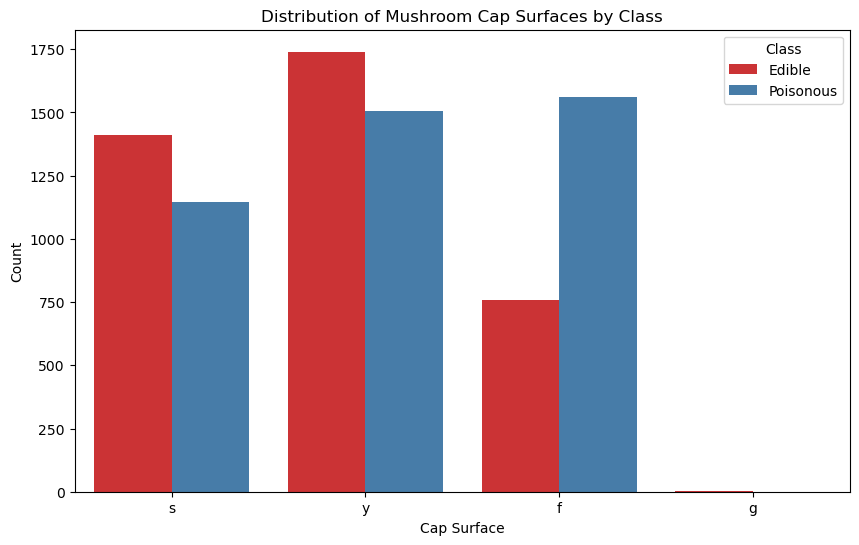
**Visualising each feature:**

* **Feature: Cap-shape**



This bar chart illustrates the distribution of mushroom cap shapes categorized by edibility. The shapes are represented on the x-axis, while the count of each shape is on the y-axis. It's evident that convex (x) and flat (f) shapes are the most common among both edible and poisonous classes, with convex shapes being slightly more prevalent in the edible category.

* **Feature: Cap-surface**



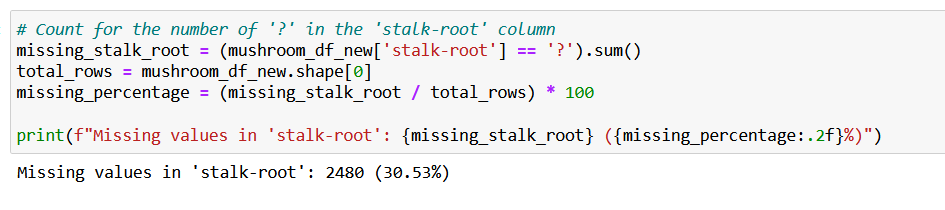
The bar chart displays the counts of different mushroom cap surfaces divided into edible and poisonous categories. Smooth (s) and scaly (y) surfaces appear more frequently among edible mushrooms, while fibrous (f) surfaces are more common in poisonous ones. Grooved (g) surfaces are the least common and appear only in the poisonous category.

These steps in the EDA process were crucial in gaining an initial understanding of the dataset, its features, and the distribution of the classes. The insights gathered from these analyses informed the subsequent stages of the project, particularly in the areas of data preprocessing and model development.

**Data Cleaning:**

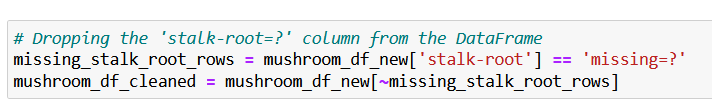
**Handling Missing Values:**

In this section, we are addressing the issue of missing values in our dataset. Missing values can hinder the performance of machine learning models, so it's essential to handle them appropriately.



We are specifically looking at the 'stalk-root' column to check how many missing values ('?') it contains. We calculate the percentage of missing values relative to the total number of rows. This helps us understand the extent of missing data in this column.

**Dropping Stalk-root: missing=? column:**



Removing the rows where the 'stalk-root' column contains the value 'missing=?'. This is a way to handle missing data by excluding the rows with missing values in this column. The resulting DataFrame, 'mushroom\_df\_cleaned,' contains these rows removed.

**Double Checking for Missing Values:**

After handling missing values, we double-check to ensure that no other missing values exist in the dataset.

there are any remaining missing values in the dataset by checking for null values. In this case, it appears that there are no other missing values in the cleaned dataset.

**Removing Duplicate Rows:**

Duplicate rows, if present, can also affect the performance of our models. We remove any duplicate rows from the cleaned dataset.

Removing any rows that are exact duplicates, ensuring that each row in the dataset is unique.

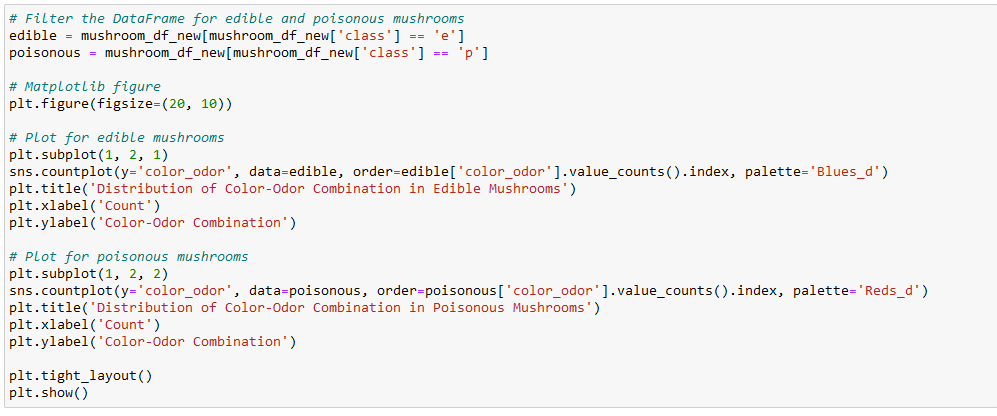
**Feature Engineering**

New features such as 'color\_odor', 'gill\_size\_attachment', and 'veil\_ring\_type' were engineered to enhance model performance. These combined attributes revealed intriguing patterns. For instance, certain color and odor combinations were predominantly associated with either edible or poisonous classes, adding valuable dimensions to the models. These engineered features provided a richer dataset, facilitating more accurate classifications.

* **Combining cap-color and odor:**

In this part, we're combining two features from our dataset, namely 'cap-color' and 'odor,' to create a new feature called 'color\_odor.' We're essentially merging the values from these two columns.

Each row in the dataset and creating a new column called 'color\_odor' that contains a combination of 'cap-color' and 'odor' values. This helps us create unique combinations representing the color and odor of each mushroom.



A comparison of a graph

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated

A graph of a bar chart

Description automatically generated with medium confidence

The graph shows a stacked bar plot of the count of mushrooms by the combination of their cap color and odor, separated into two classes: edible and poisonous. Each bar represents a unique combination of cap color and odor, with the height of the bar showing the total count of occurrences in the dataset. The green portion of the bar indicates the count of edible mushrooms with that particular color-odor combination, while the red portion indicates the count of poisonous mushrooms.

* **Combining gill-size and gill-attachment**

A screenshot of a computer code

Description automatically generated

A blue and red squares

Description automatically generated

A screen shot of a computer code

Description automatically generated

A green and red rectangular chart

Description automatically generated

The two distinct bars, each representing a different gill\_size\_attachment combination. The green section of each bar represents the count of edible mushrooms with that particular combination, while the red section shows the count of poisonous mushrooms.

* The first combination (labeled as '1' on the x-axis) has a relatively balanced distribution between edible and poisonous, with edible mushrooms being slightly more.
* The second combination (labeled as '2') is overwhelmingly represented by poisonous mushrooms, with the green section (edible) being quite minimal or even non-existent.

This kind of visualization is useful for quickly identifying which combinations of gill size and attachment are more likely to indicate whether a mushroom is edible or poisonous. For instance, the second combination might be considered a strong indicator of poisonous mushrooms, which could be useful for predictive modeling or for someone trying to identify safe mushrooms in the wild.

* **Combining veil type and ring type**

A screenshot of a computer code

Description automatically generated

A comparison of a graph

Description automatically generated with medium confidence

A screen shot of a computer code

Description automatically generated

A red square with green lines

Description automatically generated

The graph is a stacked bar chart showing the distribution of mushrooms by a combined feature of veil type and ring type, separated into two classes: edible (green portion of the bars) and poisonous (red portion of the bars). Each bar on the x-axis represents a unique combination of veil and ring types. The height of the bars indicates the total count of mushrooms for each combination within the dataset.

* The combinations represented on the x-axis are labeled as "p1\_p2", "p2\_p1", "p1\_p3", etc. These labels are likely placeholders and should correspond to the actual combined categories of veil and ring types in your dataset.
* The tallest bar, labeled "p2\_p3", suggests that this combination of veil and ring types has the highest occurrence in the dataset, with more edible mushrooms than poisonous ones.
* The bar labeled "p1\_p2" also shows a significant number of mushrooms, but in this case, the majority are poisonous.
* The bar "p2\_p1" has a lower total count, but it is composed entirely of poisonous mushrooms, indicating that this particular combination may be a strong predictor of a mushroom being poisonous.

It's important to note that the labeling of the bars (e.g., "p1\_p2", "p2\_p1") needs to be clear to draw accurate conclusions. Without meaningful labels, interpreting the graph can be challenging.

**Machine Learning:**  
- Logistic Regression:

Logistic Regression is like a straightforward tool that works well when you want to split things into two groups, like telling if a mushroom is edible or poisonous. It's like a solid starting point when you're trying to make this kind of decision. And in this case, it got everything right, like a perfect score of 1.0, meaning it could tell the difference perfectly.

Logistic Regression Accuracy: 1.0

Logistic Regression Classification Report:

precision recall f1-score support

0 1.00 1.00 1.00 664

1 1.00 1.00 1.00 636

accuracy 1.00 1300

macro avg 1.00 1.00 1.00 1300

weighted avg 1.00 1.00 1.00 1300

* Decision Trees:

Decision Trees are cool because they're simple to understand and can figure out tricky patterns in the data, even when things get a bit complicated. Just like with Logistic Regression, it nailed it with a perfect score of 1.0.

Decision Trees Accuracy: 1.0

Decision Trees Classification Report:

precision recall f1-score support

0 1.00 1.00 1.00 664

1 1.00 1.00 1.00 636

accuracy 1.00 1300

macro avg 1.00 1.00 1.00 1300

weighted avg 1.00 1.00 1.00 1300

* Random Forest:

Random Forest is like a bunch of Decision Trees working together as a team. They can handle tough situations and figure out which mushroom features matter the most. Again, it got a perfect score of 1.0, just like the other models.

Random Forest Accuracy: 1.0

Random Forest Classification Report:

precision recall f1-score support

0 1.00 1.00 1.00 664

1 1.00 1.00 1.00 636

accuracy 1.00 1300

macro avg 1.00 1.00 1.00 1300

weighted avg 1.00 1.00 1.00 1300

* Support Vector Machines Model

Support Vector Machines are a bit more complex, but they're really good at finding the perfect line between different groups. In your case, it was super close to perfect with a score of about 0.998, meaning it could almost perfectly tell edible from poisonous.

* Gradient Boosting model

Gradient Boosting is like a model that learns from its mistakes and keeps getting better. It's great at making accurate predictions, and, as expected, it got a perfect score of 1.0, just like the others.

Gradient Boosting Accuracy: 1.0

Gradient Boosting Classification Report:

precision recall f1-score support

0 1.00 1.00 1.00 664

1 1.00 1.00 1.00 636

accuracy 1.00 1300

macro avg 1.00 1.00 1.00 1300

weighted avg 1.00 1.00 1.00 1300

I picked these models because they're good at handling the kind of problem you have - deciding if a mushroom is safe to eat or not. They're also pretty smart at finding tricky patterns in the data. And the results show that they did an excellent job on your mushroom dataset.

* Contingency table:  
  Contingency Table - Model Accuracies

Accuracy

Logistic Regression 1.000

Decision Trees 1.000

Random Forest 1.000

Support Vector Machines (SVM) 0.998

Gradient Boosting 1.000

A graph showing different colors of rectangles

Description automatically generated

**Conclusions and Insights**

**Model Selection**

Given the high accuracy across all models, the choice of model could be based on factors like interpretability and computational efficiency. Decision Trees and Random Forests are recommended for their balance between accuracy and ease of understanding. These models provide insights into feature importance, beneficial for practical applications.

**Key Features**

Odor, gill color, and cap color emerged as crucial in determining mushroom edibility. These findings align with mycological knowledge, where certain odors and colors are strong indicators of toxicity or edibility. Understanding these features is crucial for anyone engaged in mushroom foraging or mycological studies.

**General Observations**

The project highlighted the complexity and diversity in mushroom characteristics. It also emphasized the importance of accurate classification, given the potential health risks associated with poisonous mushrooms.

**Future Work**

**Model Improvement**

Future work could explore more complex models like deep learning, which might capture more nuanced patterns in the data. Incorporating additional data points or experimenting with different feature engineering techniques could further enhance model performance.

**Additional Data**

Incorporating data on mushroom habitats, genetic information, or even geographic distribution could provide deeper insights and improve model robustness.

**Real-world Application**

These models could be integrated into mobile applications for mushroom foragers, providing real-time, AI-driven guidance on edibility. They could also assist mycologists in research, contributing to the broader understanding of fungi.

**Conclusion**

This project successfully demonstrated the application of machine learning in the field of mycology, particularly in classifying mushrooms as edible or poisonous. It not only showcased the power of data science in biological classification but also underscored the potential for practical, real-world applications. This work contributes to the broader field of data science by highlighting its applicability in diverse fields, including biology and environmental sciences.